DOCUMENT RESUME

ED 401 750 FL 024 247

AUTHOR	Wigglesworth, Gillian, Ed.; Elder, Catherine, Ed.
TITLE	The Language Testing Cycle: From Inception to Washback. Series S, Number 13.
INSTITUTION	Applied Linguistics Association of Australia, Clayton.
REPORT NO	ISSN-0817-9514
PUB DATE	96
NOTE	247p.
AVAILABLE FROM	Linguistics, The Faculties, Australian National University, Canberra, Australian Capital Territory 0200, Australia.
PUB TYPE	Collected Works - General (020)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	*Applied Linguistics; English (Second Language); Foreign Countries; Hindi; Language Research; Languages for Special Purposes; *Language Tests; Second Language Instruction; Test Construction; *Testing; *Testing Problems; Test Interpretation; Test Use; Time Factors (Learning); Uncommonly Taught Languages
IDENTIFIERS	Australia; Rasch Scaled Scores; Teaching to the Test

ABSTRACT

A selection of essays on language testing includes: "Perspectives on the Testing Cycle: Setting the Scene" (Catherine Elder, Gillian Wigglesworth); "The Politicisation of English: The Case of the STEP Test and the Chinese Students" (Lesleyanne Hawthorne); "Developing Language Tests for Specific Populations" (Rosemary Baker); "Developing Rating Scales for the Assessment of Second Language Performance" (Carolyn E. Turner, John A. Upshur); "Administering a Performance Test in Fiji Hindi" (Nikhat Shameen, John Read); "Specific-Purpose Language Performance Tests: Task and Interaction" (Tom Lumley, Annie Brown); "Rasch Scaling: How Valid Is It as the Basis for Content-Referenced Descriptors of Test Performance?" (Joy McQueen); "Test Interpretation, Test Use, and Pedagogical Implications" (Micheline Chalhoub-Deville); and "Investigating Washback in Japanese EFL Classrooms: Problems of Methodology" (Yoshinori Watanabe). (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

APPLIED LINGUISTICS ASSOCIATION OF AUSTRALIA

ED 401 750

The Language Testing Cycle: From Inception to Washback

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

J DUNCAN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Edited by
Gillian Wigglesworth
and Catherine Elder

BEST COPY AVAILABLE

Series S Number 13 1996

EDITORIAL COMMITTEE:

Editor: Anthony Liddicoat
Department of Linguistics
Australian National University
Canberra ACT 0200

Associate Editor: Ursula Nixon
TESOL Centre
University of Canberra

Reviews Editor: Timothy Shopen
Department of Linguistics
Australian National University

Editorial Board: Edith Bavin
Camilla Bettoni
Christopher Candlin
Frances Christie
Michael Clyne
Susanne Döpke
Marietta Elliott
Michael Halliday
Linda Hand
Debra Hoven
Mariko Iwashita
Louise Jansen
Ian Malcolm
Mike McCausland
Tim McNamara
Helen Moore
Howard Nicholas
Anne Pauwels

The *Australian Review of Applied Linguistics* is the journal of the Applied Linguistics Association of Australia, and is published twice each year. Articles on any topic of relevance may be submitted to the editor at any time. Such articles should follow the conventions set out in the style sheet in this edition. Enquiries regarding advertising in the journal should be directed to the editor.

Printed by ANU Printing Services

APPLIED LINGUISTICS ASSOCIATION OF AUSTRALIA

The language testing cycle: from inception to washback

Series S Number 13

1996

Editor of Publications

Anthony Liddicoat
Department of Linguistics
Australian National University
Canberra ACT 0200

Editors of This Volume

Gillian Wigglesworth
Macquarie University
Catherine Elder
The University of Melbourne

ISSN 0817-9514
© ALAA

CONTENTS

Page

PERSPECTIVES ON THE TESTING CYCLE: SETTING THE SCENE 1
Catherine Elder and Gillian Wigglesworth

THE POLITICISATION OF ENGLISH: THE CASE OF THE step TEST AND THE CHINESE STUDENTS 13
Lesleyanne Hawthorne

DEVELOPING LANGUAGE TESTS FOR SPECIFIC POPULATIONS ... 33
Rosemary Baker

DEVELOPING RATING SCALES FOR THE ASSESSMENT OF SECOND LANGUAGE PERFORMANCE. 55
Carolyn E. Turner and John A. Upshur

ADMINISTERING A PERFORMANCE TEST IN FIJI HINDI 80
Nikhat Shameem and John Read

SPECIFIC-PURPOSE LANGUAGE PERFORMANCE TESTS: TASK AND INTERACTION 105
Tom Lumley and Annie Brown

RASCH SCALING: HOW VALID IS IT AS THE BASIS FOR CONTENT-REFERENCED DESCRIPTORS OF TEST PERFORMANCE? 137
Joy McQueen

TEST INTERPRETATION, TEST USE, AND PEDAGOGICAL IMPLICATIONS 188
Micheline Chalhoub-Deville

INVESTIGATING WASHBACK IN JAPANESE EFL CLASSROOMS; PROBLEMS OF METHODOLOGY 208
Yoshinori Watanabe

AUTHOR NOTES 240

PERSPECTIVES ON THE TESTING CYCLE: SETTING THE SCENE

Catherine Elder
University of Melbourne
Gillian Wigglesworth
Macquarie University

The assessment of the language skills of second language learners raises complex issues which include fundamental questions about the nature of language proficiency, the validity of assessment instruments, the reliability of scores, and the manner in which the whole process may influence the curriculum. The range of issues raised are not only of importance to those who design and develop language tests; they are also of importance to a variety of people who may be involved at various points in the cycle: the funding bodies, the test takers and language teachers. The rationale behind this volume was to collect together a set of papers which would be accessible to all who might be interested in the issues and questions which are raised through the test process.

Traditionally, language testing research has been concerned primarily with an investigation of the psychometric properties of test instruments as a means of exploring their reliability and validity. Tests properties were often considered in isolation from the context in which they were implemented and the range of approaches adopted for test validation were somewhat limited. In recent years approaches to validation have broadened in scope (Hamp-Lyons and Lynch forthcoming) and the methodologies used for language testing research have become more varied with a considerable volume of studies addressing research issues in test design, delivery and assessment from more qualitative points of view. In this respect, a number of recent papers have examined a broad range of issues using qualitative methodologies or a combination of quantitative and qualitative techniques; for example: the nature of interviewer behaviour in the test environment (Lazaraton 1996, Ross 1992, Ross and Berwick 1992), the effect on discourse of different test conditions (O'Loughlin 1995, Wigglesworth 1994), the

ARAL Series S Nº 13 (1996) 1-12

relationship between discourse and rater behaviour (Douglas 1994), the relationship between the statistical properties of test items and what a close scrutiny of the items reveals about their content (Pierce 1992, Elder forthcoming), the investigation of cognitive processes of both test takers (Gordon and Hanauer 1995, Wijgh 1996) and raters (Weigle 1994). Furthermore, McNamara (forthcoming) discusses the potential of ethnomodology to illuminate the nature of the test encounter and proposes a research agenda which takes a social perspective on interaction.

This volume brings together a collection of papers in which the authors were asked to focus on one specific aspect of the testing cycle in such a way that the major stages of the cycle are each addressed. In addition, the authors were invited to approach their question by using research techniques and methodologies which did not rely exclusively on the psychometric properties of the test, and in which the analysis of the results did not rely entirely on interpretation through inferential statistical techniques although in some cases this qualitative enquiry is combined with and complements the more traditional psychometric analyses which have thus far been dominant in the field of language testing. Thus in this volume we attempt to identify a range of qualitative research techniques and to show how these may be used to address questions at every stage of the testing cycle. Finally, in documenting the whole cycle of testing (from inception to washback) we are attempting to demonstrate on the one hand the complex and multifaceted nature of the testing enterprise and on the other to show that tests do not exist in a vacuum but rather that they both shape and are shaped by the social context from which they arise.

Designed as they are around this theme, the papers in this volume cover the key stages in the cycle, from the test's inception, when the need for the test is identified and its purpose is conceptualized, through to the final stage of the testing process: namely a test's 'washback' effect (or presumed effect) on the teaching situation in classrooms which prepare students for the test. This effect may, in fact, in some cases lead back to the beginning of the cycle: that is to the reconceptualization and redesign of the test in order to bring about educational innovation. Other parts of the cycle covered in this volume are test design, test development, test

administration, the rating of test-taker performance and the subsequent interpretation and reporting of test scores.

The contributors to the volume come from Australia (Baker, Hawthorne, Lumley and Brown, McQueen), New Zealand (Shameem and Read), the United States (Chalhoub Deville), Canada (Turner and Upshur) and Japan (Watanabe). The tests they refer to represent a range of languages - English, Chinese, Arabic, Fijian Hindi. They also represent a variety of purposes, some of which seldom covered in the testing literature, ranging from the measurement of language maintenance to the assessment of language impairment.

The first paper in the volume is concerned with the political and social factors which lead to a test's inception and which determine the way in which it is used. Lesleyanne Hawthorne in 'The politicisation of English: the case of the *step* test and Chinese students' examines a specific test, the *step* test, which has been used as determiner of the residential status of immigrants from the People's Republic of China seeking asylum in Australia in the wake of the Tiananmen Square uprising. It was commissioned by the Australian government as a way of dealing speedily and economically with the huge backlog of applications on the part of those seeking official recognition of their refugee status. Thus, rather than as serving the traditional gate keeping function of barring certain people from entry, the test's prime function was to allow people to stay on in the country. In keeping with the test's political function, the setting of a language threshold within the reach of the majority of test takers was crucial. Hawthorne's account of this and other tests designed to test the language skills of intending immigrants to Australia highlights the political nature of the testing enterprise and the importance, in determining a test's fairness and validity, of considering the broader policy context from which a test arises. Issues of this kind are becoming increasingly topical in language testing circles (see, for example, Spolsky 1995, Davies 1996).

Test design issues are the focus of Rosemary Baker's paper 'Designing language tests for specific populations'. She focuses in particular on the stage in the testing cycle which precedes the actual development of test tasks or items, namely the evaluation of the suitability

of existing tests and the process of carrying out a needs analysis and of sampling relevant language skills. Baker's interest is in the measurement of impairment to language processing and communicative abilities amongst elderly bilinguals who have suffered as stroke or who are afflicted with Alzheimer's dementia. Discussions of the suitability or limitations of existing instruments are relatively rare in the testing literature in spite of this being a crucial starting point in all test development. Baker's account of the process of conducting a needs analysis is also useful, because it suggests procedures which can be followed in situations where direct observation is impracticable (in this instance the involvement of significant others and the use of questionnaires to gather information about relevant instances of language use).

The third paper by Carolyn Turner and John Upshur deals with a further aspect of test design/development which is not covered in the Baker paper, namely the development of rating scales and scoring procedures, a stage of test development which follows the selection and development of test tasks. In their paper 'Developing rating scales for the assessment of second language performance' the authors point out that rating scales are often based on a priori normative descriptions of ideal performance rather than on actual performance samples gathered in particular testing contexts and hence have dubious validity. Furthermore, they are generally constructed in such a way that assessors are required to assign very complex multifaceted performances to a single point on a scale, and the need to simultaneously consider all these different aspects of performance and to decide which ones are more salient may jeopardize the reliability of their judgments. To overcome the limitations of standard rating procedures, the authors describe the process of developing, in conjunction with primary teachers of English as a second language, an empirically derived scale which is set up in such a way that raters are limited to making a series of simple binary choices based on the presence or absence of only those features of performance that define the boundaries between score levels. This method was found to be both efficient and reliable and, since it is based on what teachers 'value and are able to recognize', it is can be usefully applied to the assessment of achievement in other educational contexts.

Nikhat Shameem and John Read are concerned less with test design (although details of task content and of rating scales are included as background information) than with test administration. Having designed a test to measure proficiency in a preliterate vernacular language spoken by a group of immigrants to New Zealand, 'Administering a performance test in Fiji Hindi' addresses the practicalities of administering such a test in a situation where there is no built-in institutional motivation to sit the test, where the emphasis is on spontaneous language use (rather than on formal knowledge), where the interlocutor is known to many of the candidates, and where there are no established native speaker norms and moreover few 'fully fledged' native speakers in a position to interpret such norms.

The paper suggests an important role for language testing in documenting language attrition and language maintenance in multilingual settings, a role which is seldom exploited. Surveys of language loss and maintenance are all too often based on self-report data, the validity and reliability of which will depend on the way in which questions are formulated and on respondents' level of awareness and commitment to the enquiry.

While this paper deals with a very particular population of test-takers (and this is true also for the Baker paper), we feel that the issues raised are generalisable to other communicative testing situations in that what they all attempt to capture is 'real world' language use with all the idiolectal variation that this entails. Classroom-based assessment frequently involves simulations of language use in domestic settings, assessment tasks are often carried out by a familiar person (e.g. the language teacher) and may be assessed by non-native or 'background' speakers, who like the raters in the Shameem and Read study, may not have full control of the target language and/or (particularly when dealing with learners who speak different dialects of the target language) may be faced with difficult decisions about which norms to apply. As Shameem and Read point out, the quest for validity can be at odds with the requirements of reliability and practicality, and all testing involves a trade-off between these competing demands.

The tension between validity and reliability re-emerges as an issue in 'Specific Purpose language Performance tests: task and interaction'

by Tom Lumley and Annie Brown. The research they describe draws on data from the speaking component of an occupation specific test of English measuring the English language skills required by overseas-educated nurses in order to effectively fulfil their professional duties in an Australian medical context.

The paper is concerned with what a qualitative analysis undertaken after a test has been administered can tell us about a test's content validity and about the reliability of the test elicitation procedure. Feedback is gathered from occupational experts (in this case members of the nursing profession) on the test materials and on audio taped samples of test interaction in order to establish the authenticity of test tasks in relation to the real domain which they are intended to reflect (i.e. the nurse-patient encounter). Discourse analysis is also undertaken to identify the extent to which interlocutors identify with the role prescribed for them by the test developers and the features of their behaviour which are likely to make the candidates' task easier or more difficult. The authors conclude that while the quest for task authenticity it is desirable, it is inevitably constrained by the requirements of a test which necessitate some degree of standardization across interviews. In the interests of fairness to candidates, a balance needs to be struck between reducing variability by refining the test prompts and by training test interlocutors to comply with set procedures on the one hand, and, on the other, allowing the dynamics of communicative interaction to be played out in as natural a manner as possible. In similar vein Suchman and Jordan point to the dangers of imposing too many constraints on the interview situation:

> Stability of meaning, the real basis for standardization and ultimately for validity, requires the full resources of conversational interaction.
>
> (Suchman and Jordan 1990: 233)

Joy McQueen's paper 'Rasch scaling: how valid is it as the basis for content-referenced descriptors of test performance?' focuses on a later stage of the testing cycle, namely the reporting of test performance. She considers the validity of performance descriptors at three levels of ability derived from a post-hoc content analysis of test items undertaken by test developers on a multiple-choice test of beginning reading proficiency in

Chinese as a Foreign Language. Again, following the approach adopted by Turner and Upshur (see above), the emphasis is on producing empirically-derived scales rather than basing the descriptors on some a priori notion of what performance should look like.

McQueen's approach to this validation exercise is a novel one which will be of interest to other researchers. She surveys the research literature on second language reading and vocabulary development and on test method facets (Bachman 1990) and identifies a number of factors relating to item content and structure which might be expected to predict the rank order of item difficulty as mapped out by a Rasch analysis. Correlations between these factors and item difficulty values are performed and the performance descriptors (which have been drawn up independently by the test developers) are then examined to determine whether those features which are found to be effective predictors of item difficulty are adequately represented in the wording of the descriptors. She finds that the mapping procedure is in general a valid approach but that the quality of the descriptors could be improved with closer attention to insights gathered from the research literature. It may be considered that such attention to detail (or indeed to theory) is not warranted in this case, since the function of these descriptors has more to do with face validity (i.e. public recognition of early achievements in LOTE learning) than with construct validity and since the length of the descriptors is severely constrained by the limited space available on the certificates issued to candidates. Nevertheless, the methodology proposed by McQueen has great potential in more high stakes testing situations, where a lot hinges on test outcomes and on the inferences which are drawn from them. A further application of this approach would be in the formulation of profiles of performance which describe learner performance 'standards' or 'benchmarks' or 'or 'competencies' at different levels of achievement and are currently being imposed nation-wide in Australia and other countries for educational accountability purposes. A common criticism of these outcome statements has been that they are produced in a 'top down' fashion and have no basis in reality (Brindley 1995, Moore 1996). The cross-referencing of theoretical insights to empirically derived analyses of item behaviour would therefore appear to hold considerable promise.

'Test Interpretation, Test Use and Pedagogical Interpretations' by Micheline Chalhoub-Deville deals with the interpretation of test scores and the potential uses which can be made of information provided by candidates' performance on a test of Arabic as a foreign language (AFL). The author uses speech sample analyses to explicate the meaning of the dimensions that emerge from a quantitative analysis of test performance and considers how candidates' weaknesses (e.g. grammatical inaccuracy, lack of creativity, unwillingness to supply detail) as revealed in their test performance, could be addressed in the AFL classroom. She thereby points to the close connection between the test and the pedagogical context in which it is administered. She sees the link between the test and the classroom as a two way process: teachers should provide candidates with communicative learning experiences which will allow them to do well on performance tests such as the one she describes, and the scores assigned to test performance should be derived from criteria nominated by end-users (in this case teachers) so that their meaning is clear and so that appropriate measures can be undertaken to improve candidate performance. Thus this paper raises the important and controversial issue of the relationship between language proficiency testing and classroom practice and stresses the importance of collaboration between test developers and ender-users (a view which is shared by Lynch and Davidson, 1994 and Alderson and Clapham 1995). While Chalhoub-Deville's focus here in on teachers and testefs of Arabic as a Second Language, the points she raises have implications beyond that immediate group.

The relationship between testing and classroom practice is further investigated in the final paper 'Investigating washback in Japanese EFL classrooms; problems of methodology' by Yoshinori Watanabe who writes about highstakes EFL tests adminstered at the end of secondary school as a screening device for university admission in Japan. Watanabe extends the scope of language testing research beyond the traditional boundaries, i.e. of construct and content and criterion-related validity into the social consequences of test use, by considering the impact of university entrance tests on teacher and learner behaviour in different types of classrooms. In this paper he is concerned primarily with

methodology and with the insights which an ethnographic approach can offer those interested in undertaking this kind of research. The importance of washback and its relationship to a test's 'consequential validity' has been acknowledged by a number of researchers (e.g. Alderson and Wall 1993 Frederickson and Collins 1989, Linn 1989, Messick 1989, Moss 1992, Gipps 1994) and a forthcoming issue of the journal Language Testing (13,3) is devoted entirely to this question. However, the phenomenon is thus far under researched perhaps because, as Watanabe's research seems to indicate, it is extraordinarily difficult to establish a causal relationship between the test and what happens in the classroom given the large number of variables involved.

It is hoped that non testing specialists will find this volume accessible and interesting enough to dispel the widely held prejudice that testing is a purely technical activity. Language testing, as Alan Davies has proposed, can be located at the centre of applied linguistics

> because it operationalises hypotheses and provides practical scenarios.... Language testing firmly tackles such concepts as native speaker, level of proficiency, criterion, test, language, all no man's land concepts in applied linguistic discussions and gives them definition and operational effect. By so doing language testing acts both as research methodology for applied linguistics and as data for model extension and building.
>
> (Davies 1990: 70, 83)

Referring to the content of the papers in this volume we would add to Davies' list of the concepts defined, elaborated and/or operationalised by language testing that of language variety, language maintenance, language impairment, language policy, spoken interaction, second language reading and decision-making behaviour.

While the test encounter imposes constraints on language behaviour which are related to the particular purpose of the entreprise (and for this it is often criticized), it also provides us with a well-specified context within which this behaviour can be analysed and reveals factors that we might be otherwise be unaware of governing communication in a range of interpersonal and institutional contexts. The implications of language tests, as is evidenced in a number of the papers in this volume, extend beyond the test encounter itself and it is important that both test

developers and test consumers are mindful of both the power and the limitations of the testing enterprise.

REFERENCES

Alderson, C. and C. Clapham (1995) Assessing student performance in the ESL classroom. *TESOL Quarterly* 29: 184-187.

Alderson, J. C. and D. Wall (1993)Does washback exist? *Applied Linguistics* 14, 115-129.

Bachman, L. (1990) *Fundamental considerations in language testing.* Oxford, Oxford University Press.

Brindley, G. (1995) Assessment and reporting in language learning programs: purposes, problems and pitfalls. Paper presented at the International Conference on Testing and Evaluation in Second Language Education, Hong Kong University of Science and Technology, June 1995.

Cumming A. and R. Berwick (1996) *Validation in language testing.* Clevedon, Avons, Multilingual Matters.

Davies, A. (1990) *Principles of language testing.* Oxford, Basil Blackwell.

Davies, A. (1996) Migrant gatekeeping through English proficiency is efficiency replacing equity. Paper presented at the Language Testing Research Colloquium, Tampere, Finland.

Douglas, D. (1994) Quantity and quality in speaking test performance. *Language Testing* 11,2: 125-144.

Elder, C. (forthcoming) The effect of language background on 'foreign' language test performance: the case of Chinese, Modern Greek and Italian. *Language Learning* 46, 2.

Frederiksen, J. and A. Collins (1989) A systems approach to educational testing. *Educational Researcher* 18,9: 27-32.

Gipps, C. P. (1994) *Beyond testing.* Falmer Press.

Gordon, C. and D. Hanauer (1995) The interaction between task and meaning construction in EFL reading comprehension tests. *TESOL Quarterly* 29: 299-324.

Hamp-Lyons, L. and B. Lynch (forthcoming) Perspectives on validity: A historical analysis of the LTRC. In Kunnan A. (ed.) *Issues in language testing research: From conventional validation to fairness.* New Jersey, Lawrence Erlbaum.

Lazaraton, A. (1996) Interlocutor support in oral proficiency interviews: the case of CASE. *Language Testing* 13,2: 151-172.

Linn, R. K. (1989) (ed.) *Educational measurement.* (3rd edition). American Council on Education, Washington Macmillan.

Lynch, B.K. and F. Davidson (1994). Criterion-referenced language test development: Linking Curricula, Teachers, and Tests. *TESOL Quarterly* 28,4: 727-743.

McNamara, T. (forthcoming) Interaction in second language performance assessment. *Applied Linguistics.*

Messick S. (1989) Validity. In R. Linn (ed.) *Educational measurement* (3rd edition).

Moore, H. (1996) Telling what is real: competing views in ESL assessment. *Linguistics in Education* 6, 8: 189-228.

Moss, P. A. (1992) Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research* 62, 3: 229-58.

O'Loughlin, K. (1995): Lexical density in candidate output on direct and semi direct versions of an oral proficiency test. *Language Testing* 12, 2: 217-237.

Pierce, B. (1992) Demistifying the TOEFL reading test. *TESOL Quarterly* 26: 665-691.

Ross, S. (1992) Accommodation questions in oral proficiency interviews. *Language Testing* 9: 173-186.

Ross, S. and R.Berwick (1992) The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition* 14: 159-176.

Spolsky, B. (1995) *Measured words: The development of objective language testing.* Oxford, Oxford University Press.

Suchman, L. and B. Jordan (1990) Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association* 85, 409: 233-241.

Weigle, S. (1994) Effects of training on raters of ESL compositions. *Language Testing* 11, 2: 197-223.

Wigglesworth, G. (1994) The effect of information gap on elicited discourse of candidates in an oral interaction test of English. Paper presented at the Language Testing Research Colloquium, Washington DC, March 4-7.

Wijgh, I. (1996) A comunicative test in analysis: Strategies in reading authentic texts. In Cumming, A. and R. Berwick (eds) *Validation in language testing*. Clevedon Avon, Multilingual Matters.

THE POLITICISATION OF ENGLISH: THE CASE OF THE *step* TEST AND THE CHINESE STUDENTS

Lesleyanne Hawthorne
Bureau of Immigration, Multicultural and Population Research

ABSTRACT

The past decade in Australia has been characterised by a striking growth in the application of language testing, frequently in contexts governed by macro-political pressures. Despite this, the literature to date has rarely placed its first emphasis on the pragmatic considerations which typically give rise to test development, and impact on test design as well as administration outcomes. To illustrate the significance of this, this chapter explores the recent pressures surrounding the evolution and implementation of the Special Test of English Proficiency (*step* test) - a form of ESL testing designed not merely to assess linguistic competence, but to play a central role in the determination of residential status for substantial numbers of asylum seekers.

INTRODUCTION

> The English language test being taken in capital cities this week by more than 7,000 people claiming refugee status makes for an unusual event. The test has been the subject of complaints from some of those taking it and has caused unease among some Australians concerned about its fairness. Yet it is hard to see how either of these concerns is justified. The test is a reasonable part of a necessary administrative response to a complicated situation which has to be resolved.
>
> (Sydney Morning Herald, 1994a)

Australia can claim a long-established precedent in the use of language testing for political purposes - most notoriously through administration of the 'dictation test' as a central tool of the 1901 Immigration Restriction Act[1]. Masquerading as an examination of literacy, the aim of this test was in fact explicit racial exclusion. Applied at the discretion of federal immigration officers, it originally required the correct transcription of 50 words in any European language, later extended to 'any prescribed' language. Within the period 1902 to 1946, rigorous application of the test

ARAL Series S Nº 13 (1996) 13-32

ensured only around 125,000 members of 'the alien races' were admitted to Australia - the vast majority of these former domiciles, or others awarded 'Certificates of (Test) Exemption' (York 1993: 6). According to the 1922 Commonwealth Year Book:

> In pursuance of the 'White Australia' policy, the general practice is not to permit Asiatics or coloured immigrants to enter Australia for any purpose of settling permanently. Exclusion is effected, where necessary, by the application of a dictation test... In other words, the test when applied is an absolute bar to admission.
>
> (York 1993: 4)

The past decade in Australia has similarly been characterised by a striking growth in the application of language testing, frequently in contexts governed by macro-political pressures (see eg McNamara 1990, Hawthorne 1994, 1995). Two examples serve to illustrate this trend in relation to English. The Occupational English Test (OET), developed by McNamara in the late 1980s to screen the linguistic competence of overseas qualified health professionals, in fact operates as a powerful gatekeeper to both the opportunity to migrate (within the points tested skill categories) and to Australian professional registration. In terms of nurses, for instance, a recent test administration at a prominent overseas post led to the rejection of a substantial majority of candidates through high OET failure rates (Hawthorne & Toth forthcoming). In relation to overseas qualified doctors, a pass in the OET has been until recently an essential pre-requisite to even sitting the Multiple Choice Question and Clinical examinations of medical knowledge - with the potential to delay candidates for years should they have difficulty in passing the speaking, listening, reading or writing tests.

A second example of what I have previously termed 'the politicisation of English' relates to the introduction of the Australian Assessment of Communicative English Skills (access:) test. This mandatory English language test was designed in the early 1990's to assess the language level of potential skilled migrants at Australian overseas posts - with passes essential in 114 prominent occupations defined as 'Occupations Requiring English' (ORE). This testing represented a radical departure from past practice in a number of ways. Firstly, the allocation of a maximum 20 points to Independent candidates possessing 'vocational' ESL competence offered a substantial

competitive advantage to candidates from English-speaking countries, or from countries with strong traditions of teaching English (such as Hong Kong or India). Secondly, the access: test pass mark was set by the Department of Immigration and Ethnic Affairs rather than test developers - a situation which allowed the assessment of English to become a direct and malleable instrument for the control of skilled immigration intakes. Thirdly, 'Occupations Requiring English' embraced an extremely comprehensive list of professions and trades, spanning health, engineering, education, public safety, policing, electrical trades etc (Commonwealth of Australia Gazette 1993). ORE Concessional as well as Independent principal applicants who failed to demonstrate 'vocational' English would henceforth not be permitted to migrate to Australia - however high the number of points they scored on other criteria.

Despite the growing emphasis placed on language testing over the past decade, the Australian literature to date has rarely placed its first emphasis on the political and pragmatic considerations which typically give rise to test development - impacting on test design as well as administration and outcomes. To illustrate the significance of this further, this chapter explores the recent pressures surrounding the evolution and implementation of the *step* test ('Special Test of English Proficiency') - a form of ESL testing designed not merely to assess linguistic competence but to play a central role in the determination of residential status for substantial numbers of asylum seekers.

In so doing the *step* test represents a fascinating contrast in intent as well as outcomes to the notorious 'dictation test' operating earlier this century. Designed to appear as a gatekeeping device, *step* has in fact been constructed to facilitate the Australian Government's acceptance of a large majority of asylum seekers - imposing a judicious appearance of control over what was potentially an unmanageable, expensive and diplomatically awkward situation. In a neat historical irony, both the dictation and the *step* tests have primarily been structured to screen applicants for permanent residence from (the People's Republic of) China. The transformation in design, as well as intent of the testing applied to them, symbolises the profound shift which has occurred in the nature of Australia's immigration program, together with the shift in Australia's perception of the value of Asia.

THE POLITICAL CONTEXT OF THE step TEST

The circumstances which gave rise to the development of the *step* test are integrally linked to the dramatic expansion of Australia's 'export education' program which occurred throughout the 1980s - in particular, the growth in English Language Intensive Courses for Overseas Students (the ELICOS sector) towards the end of the decade.

According to Australian Bureau of Statistics (ABS) data, the percentage of short-term arrivals stating education to be their prime purpose of visit surged from 30-40 per cent of total arrivals in the 1960s and 1970s, to 60-70% within the past decade (Shu & Hawthorne 1996). Asian countries have always been the dominant source of Australia's international students, despite acceptance of people from more than 80 different countries. However, numbers increased markedly after the mid-80s liberalisation of student entry requirements - rising by 250 per cent, in particular through the admission of ELICOS students from the People's Republic of China (PRC). Although this trend was followed from 1989 by progressive tightening of entry criteria for selected countries, such controls came too late to prevent serious levels of overstaying (Hansard 1994).

Growth in international student numbers has in fact always had a substantial impact on Australia's immigration program - with large numbers of Malay and Indonesian students from the 1950s onwards converting to permanent resident status (Shu & Hawthorne 1996, Waldron 1992). A recent study, entitled *International Students and Immigration*, indicates this trend persists, with around 50% of surveyed students ultimately intending to migrate to Australia (Nesdale et al, 1995). The prime source of student overstayers in recent years has been the People's Republic of China - with the majority of students reaching Australia on a full-fee paying basis within the ELICOS sector. From the start it was evident that 'backdoor migration' or 'English for visa' was a prime motivation for student arrivals (Shu & Hawthorne 1996). Of the 16,500 PRC ELICOS students reaching Australia between 1987 and the events at Tiananmen Square (June 1989), only around 10 per cent ultimately departed (Davenport 1990; Birrell 1993, 1994a and b). In the years immediately following, 9,000 spouses and dependent children joined the total of more than 19,000 Chinese nationals already resident

in Australia. Prior to policy tightening, moreover, thousands of additional PRC students 'in the pipeline' continued to arrive. By October 1992, according to Australia's Federal Minister for Justice:

> About half (of all) PRC student arrivals have either overstayed, or applied for refugee status which has delayed their departure.
>
> (DIEA 1992: 2)

Rather than profiting from PRC students' fees, as expected, the Australian government had to face their abrupt metamorphosis into an on-shore quasi-refugee movement. In 1994 the Sydney Morning Herald summarised the situation as follows:

> After June 4, 1989, many thousands of Chinese continued to come to Australia as 'students'. Many were enrolled in private English-language schools. Like a great many of those who came before June 4, 1989, many were economic refugees who had managed to borrow enough money to get a place in an English-language school in Australia but had no intention of returning to China. Because of the high rate of overstaying by such 'students', controls on this category of entry have since been tightened. But before that necessary change was made, very large numbers of 'students' who had entered Australia after June 4, 1989, were claiming they should be allowed to remain - just as those who had come before June 4, 1989, had been permitted to remain under Mr Hawke's rash promise of sanctuary.[2]
>
> (Sydney Morning Herald, 1994b)

In part as a result of this, by November 1 1993 Australia found itself confronting a need to determine the status of three distinctive student groups - ultimately amounting to 49,672 persons[3] (DIEA 1996):

- PRC nationals in Australia at the time of the Tiananmen Square uprising (Class 815 entry permits);
- some better-qualified asylum seekers and nationals of Sri Lanka and the former Yugoslavia temporarily in Australia, provided they meet certain criteria (Class 816);
- some highly qualified students undertaking post graduate study in Australia (Class 818)

(DIEA 1995a:1).

Although permanent residence (in line with Hawke's promise) would be extended to virtually all Class 815 applicants, determination

of status for those subsequently reaching Australia threatened to be a dauntingly long and expensive process. According to the United Nations Protocol, to which Australia was signatory, a refugee was by definition someone outside his/her country of nationality or usual place of residence, who was:

> unable or unwilling to return or to seek the protection of that country because of well-founded fear of being persecuted for reasons of... race, religion, nationality, membership of a particular social group, or political opinion'.
>
> (DIEA 1995b:1)

Throughout the 1980s, Australia had to a degree dealt with illegal immigrants and asylum seekers by declaring a successive series of amnesties (DIEA 1988) - a viable option where there were relatively small numbers, and one essentially adopted in 1993 in relation to Class 815 applicants (DIEA 1995a). It would be politically problematic, however, to extend such largesse to all other potential asylum seekers. By 1990/91, on-shore refugee applications had reached an historic high of 16,937, with around 77% of these lodged by former PRC students - the 12,768 people plus dependants who had reached Australia after June 1989 (Birrell 1995: 48.). By March 1996, these students represented by far the largest group of Class 816 principal applicants - 10,596 out of a total of 14,274, compared with just 569 from Pakistan, the second biggest country of origin (DIEA 1996). The dominance of PRC students was even greater among Class 815 applicants, 12,857 people out of a total of 12,897 principal applicants.

By 1993 Australia's system for the determination of refugee status involved two possible stages: a preliminary assessment of applicants' claims against the UN definition by a Department of Immigration and Ethnic Affairs (DIEA) officer, followed by a review 'where people whose applications for a protection visa have been refused can appeal for an independent assessment' (DIEA 1995b:2). Within this appeal process, the Refugee Review Tribunal was required to re-analyse all written evidence - if necessary taking the case to a personal hearing. Where there was a perceived error of law, rejected applicants retained the right of further appeal to the Federal Court - a process leading to decisions being upheld, or a direction for their reassessment. Contemporary evidence suggested the majority of applicants were likely to fail, given that in 1992/3 only 4.5 per cent were successful at the primary level, together with

6.3 per cent at the review stage (Birrell 1994a: 35). Despite this, by 1993/4 around 72% of applicants refused protection visas were electing to seek a review of their case - a trend with major ramifications. According to Robert Birrell, a prominent critic of government immigration policy, 'By 1993 over 50 per cent of the Federal Court's time was being taken up by immigration cases and the proportion was still rising' (Birrell 1993: 17). The cost of processing such applications was high - amounting to $50.7 million per annum by 1994/5 (DIEA 1995b: 3).

Even discounting resource issues, ensuring a fair determination of refugee status could be extremely problematic. Though the aim of such assessment was to determine if individuals had a 'well-founded fear of persecution', no approach for verification could be made to country of origin. Rather, according to DIEA:

> Assessments of claims for refugee protection are made on an individual basis using all available and relevant information concerning the human rights' situation in the applicants' home country.
>
> (DIEA 1995b:2)

Australia faced a major dilemma in assessing the validity of claims of PRC students. The data presented in *A Country Profile: China* by the (Refugees, Immigration and Asylum Section of the Department of Foreign Affairs and Trade (DFAT)) suggests there was in fact minimal evidence that PRC students reaching Australia following the 1989 events at Tiananmen Square had any real grounds for securing protection (DFAT 1994). Key players in the pro-democracy movement had been banned from leaving China in the years immediately following 1989. Although it was presumed by the Australian government 'as a matter of course' that activists in Australia had been 'watched closely', according to DFAT:

> Until now we are not aware of any instance in which a returnee who has been determined not to be a refugee under the DORS process in Australia has returned to China and suffered persecution...Our information gives similar advice about returnees from other countries... While some have indicated they were questioned by immigration primary officers about the length of time they have been away from China, the questioning does not seem to have been serious and none was taken off the primary line for questioning.
>
> (DFAT 1994:43)

In August 1992, moreover, the Chinese government had formally encouraged overseas students to return, providing an assurance it would 'not conduct political investigation of returnees who expressed dissent while overseas' (DFAT 1994:30). This view was subsequently endorsed by independent media (eg Rubin 1995).

Despite the dearth of evidence to support their claim for asylum, by mid 1993 it was clear that the great majority of PRC students would pursue all avenues to stay in Australia - an expectation by no means shared by the majority of Australians. The government's first 'November 1 1993 Decision', guaranteeing permanent residence to PRC students resident by June 1989, was greeted with little surprise, given the commitment by former Prime Minister Hawke that none would be forcibly repatriated. The controversy lay in the accompanying statement that subsequent PRC and other student arrivals would also be eligible for permanent residence - a decision viewed by some as an unwarranted 'capitulation'. According to Birrell, this decision appeared to be based on the following reasons:

> First, the government's concern that the backlog of on-shore asylum cases was growing to the point where it could take years to clear it. Meanwhile, those awaiting processing would have accumulated a stronger claim to remain in Australia. Second, the government was worried about the willingness of the Federal and High courts to take up appeals on administrative law grounds against the judgments of its Refugee Review system. These appeals could further clog the judicial system, making it difficult to bring the backlog of asylum cases to finality.
>
> (Birrell 1994a:35)

The explanation for this decision offered by DIEA is far more positive. In total, the Class 816 category covered 18,957 persons, including 14,349 principal applicants: with 67 per cent of the total from the PRC, 3.8 per cent from Sri Lanka and 2.2 per cent from the former Yugoslavia (February 1995 figures). 74% of these people had applications for refugee status pending, amounting to 13,039 people across all three classes. Australia was thus confronting a substantial queue of comparatively well-qualified people waiting for asylum, who had already been resident a number of years, endured years of uncertainty, and in terms of human capital had clear potential to contribute to Australia's skill migration program. Skills-based assessment could afford

individual applicants a means of circumventing the lengthy and expensive process of determination of refugee status, while providing them with a far greater chance of success. From a government perspective, this approach presented a breathtakingly simple, cheap and compassionate solution - fortuitously coinciding with a drop in external demand for Australian skilled migrant places.

All outstanding applications for refugee status were frozen, pending the outcome of Class 816 assessments, with those succeeding immediately taken off the list. Although applicants would be required to meet health, character and age standards (aged less than 45 on 1 November 1993), Class 816 applicants would additionally be assessed by skill migration criteria: a proven ability 'to meet English language and post-secondary education standards' (DIEA 1995a:1). Such requirements were routinely applied to off-shore applicants, and seemed eminently fair to the Australian government - securing bipartisan support. Writing to The Age, the Minister for Immigration and Ethnic Affairs asked:

> Is it unreasonable for Australia to ask those seeking permanent residency under a skills category to meet requirements of English language ability, particularly when (many of those protesting) came here to learn English in the first place?
>
> (The Age 1994:7)

CONSTRUCTING THE step TEST

Immediately following the Minister's 1993 announcement, a DIEA 'November 1 Taskforce' was established in Canberra in December 1993, given responsibility for managing Class 815, 816 and 818 applications - including relevant policy, legislation and implementation over a two year period. One hundred and fifty staff were recruited for the Sydney and Melbourne offices. Since assessment would be applied to qualifications as well as English as a Second Language skills, immediate liaison was established with the National Office of Overseas Skills Recognition (NOOSR), which would provide initial training, in addition to taking responsibility for the detailed assessment of the more complex 'borderline' professional qualifications. The Department of Industrial Relations (DIR)[4] and state/territory trade authorities would assess trade qualifications. Within this period, the mechanics of the subsequent *step*

test were also being devised - the aim being the imposition of minimal financial burden on the Australian taxpayer. A sum of around $500,000 was allocated to the November 1 Taskforce - sufficient to cover test development, the training of raters and markers, and local as well as international trialling - a sum off-set by the $830 application fee paid by each candidate (DIEA 1994). All further costs associated with the test (for administration, marking, accommodation hire, security and appeals) would be covered by subsequent applicants' fees: $145 to sit all three test modules for the first time, $45 to take individual modules on a re-sit basis, and $45 to cover appeals in borderline cases. Applicants passing would be permitted to stay in Australia - assuming the acceptability of character, health and qualifications checks. Those failing would revert to the determination of refugee status - with any failing facing 'mandatory removal' from Australia.

Despite the rumours immediately circulating within the Chinese press, DIEA was in fact determined from the outset to encourage candidates to canvass the fullest range of possibilities to secure permanent status. According to a spokesperson:

> What we're saying is, the law is there providing an opportunity for people of a wide range of options, and people should ensure in all fairness to themselves that they look at the full gamut of possibilities... eg as spouses, Employer Nomination Scheme, Special Talents... There's nothing more difficult and frustrating after your options have closed than to find that you might have had another option earlier!...(F)rom that day onwards, we have been virtually saying 'Get on with it!'

Within weeks of the November 1 announcement, DIEA had approached the National Centre for English Language Teaching and Research (NCELTR) at Macquarie University, with a request to design and administer the functional English test which would play a central role in Class 816 assessments - the *step* test. Selected by DIEA on the basis of being a 'highly eminent group of experts', NCELTR would lead a consortium including the Language Testing and Research Centre (University of Melbourne) and the Adult Migrant English Service (NSW) - the same team which in 1992-93 had designed the well-regarded access: test for ORE skilled migrants. From the outset, according to Professor Christopher Candlin, NCELTR's Executive Director, the Department's aim in terms of the test was both rigour and fairness - building into the

process, for example, means of catering to special needs applicants ('deaf, injured, paraplegic candidates'). Given the high stakes of the test, a DIEA spokesperson stated, 'One of the key things we wanted in using NCELTR was to guarantee that not only was it a fair process, but it also had to have all the procedures and processes that showed it was a fair process.'

The *step* test would differ in two key respects from the pre-migration access: test which was designed to determine the linguistic competence of potential skilled migrants (Hawthorne 1995). Firstly, it would have a clearly defined audience who were relatively well-established in Australia, and required English 'for general lifestyle... rather than senior professional (work) or the like'. It would thus require a less exacting language level than access: ('functional' rather than 'vocational' English), with successful candidates scoring across a defined band. Secondly, procedures would be simplified by concentrating on listening, reading and writing - eliminating the more complex and costly testing of speaking.

These band descriptors, defined early in 1994, were immediately incorporated into Australian law. Exemptions from taking the test were awarded by DIEA to applicants who had successfully completed a year's full time study (or equivalent) in Australia, an accredited post-secondary course or equivalent, an overseas associate diploma in which the primary language was English, or a pass within existing accredited language tests (International English Language Testing Scheme, the Occupational English Test, or access:) within two years of Class 816 application. On the basis of this, approximately 30 per cent of applicants were exempted from the test, reducing potential *step* candidate numbers to around 9,800. These candidates were offered two opportunities to sit, with a built-in appeal procedure available for borderline failures. Meticulous care was taken in the trialling of *step* versions, to ensure the maintenance of complete test security - a vital issue in such a high-stakes process. The writing test was also to be double-rated, as a safeguard against unintentional rater inconsistency and/or bias.

In terms of standards, the Department's advice to the consortium was to be as generous as possible in the determination of cut-off levels, while maintaining the integrity of the test, within the band specified by the Minister. The timelag between the 1 November 1993 policy announcement and the test administrations in fact guaranteed candidates

12 to 18 months preparation, during which NCELTR was encouraged to provide the maximum possible information. A manual entitled *step Forward* was commissioned by NCELTR and externally produced, providing detailed information on test format as well as approximate level of difficulty. Exercises within this manual made it clear that *step* tasks were designed to screen candidates' capacity to participate in common work, social and lifestyle situations, such as shopping, use of public transport, job interviews etc. (NCELTR 1994). As part of its outreach to potential candidates, NCELTR sent a letter to known ESL providers Australia-wide, informing them of the test, and encouraging them to offer *step* preparation classes. As a consequence, documentation of test specifications was provided to 85 ESL institutions, supported by continuous feedback from the NCELTR *step* office. *step* information sessions were scheduled for Sydney and Melbourne providers, together with parallel events targeting the ethnic media and community representatives. This whole complex test administration process was managed by Jennifer McGrath on behalf of NCELTR. According to a DIEA spokesperson:

> When the regulations came out... we were encouraging people to learn like mad! Extensive private courses were opening in Sydney and Melbourne. The feedback was that there were very few English language classes that had vacant positions on the east coast! Courses at TAFE were being opened up to serve their interests. The (applicants) were very very diligent, and they went for it!

THE ASSESSMENT PROCESS

The first administration of the *step* test was heralded by an unprecedented Editorial on English language testing in the Sydney Morning Herald (3 November 1994) - a reflection of the degree to which ESL issues were now intersecting with mainstream politics:

> By November 1, 1993, there were about 20,000 people in Australia seeking refugee status. Normally it would have taken years to process that many applications. Instead of having to wait that long... those doing the test this week are being given an exceptional opportunity. It is to gain permanent residence under the new category, with its emphasis on skills - one of which is proficiency in English. Some will fail. They will then be reassessed for their

> eligibility as refugees. Again, many will be accorded refugee status. But, as long as that process is fair, they can hardly claim of unfair treatment.
>
> (Sydney Morning Herald 1994a)

By the time the *step* test administrations began, immense anxiety was being expressed within the Chinese community. Two major hurdles confronted the candidates: linguistic and vocational assessment. In relation to the former, there was intense speculation about the imposition of secret DIEA quotas; concern that the pass level would be excessively high - so high it would be beyond the ability of most applicants (Chinese Herald 1994), and not even Year 12 native speakers would be able to pass it (New Migrant 1994). Some critics claimed this pass level was being imposed as a cost-saving measure, to ensure successful applicants would later be ineligible for free tuition in the Adult Migrant English Program (The Tide 1994a). Further concerns centred on China: that significant dangers faced eventual returnees (United Chinese Times 1994a); that public administration of the *step* test would inevitably place failed candidates at a high risk of retaliation (The Age 1994); and overriding all, that use of linguistic and qualifications criteria to decide the status of asylum seekers was invalid and inhumanitarian. This view persisted despite repeated assurances from DIEA staff that failing Class 816 and 818 assessment did not entail a risk of immediate deportation from Australia - rather a resumption of the interrupted process of determination of refugee status.

In Candlin's view, the early *step* administrations went some way to reassuring candidates:

> I can well remember a test administration in Western Sydney in which... the testees stood up and clapped! Not an experience I've ever had before, I have to say... The feedback we've had from candidates and the teachers of *step* preparation classes has been that the test has had a very positive effect. We've had people saying they've been 'forced' to undertake English classes, and thus had improved their language skills... (Community consultations show) that people had, as a result of taking the *step* preparation classes for the test, got themselves locked into a scheme of taking further study and further courses... People had somehow been shaken into the possibilities of further training.

This view is endorsed by a Department spokesperson:

> The stories were... absolutely fascinating! The way it bonded with the host community was wonderful... You'd have Australians ringing up saying 'I have this young lady (or man). She comes to my local church, and we've been working feverishly every night to get her English up to scratch! Are there any other books we can read?... The results I hope will reflect that kind of dedication... The idea is to be inclusive, not exclusive... If more people can get to this level, the happier we are! The ultimate at one end is if 100 per cent pass, we would be over the moon. Because (the test) would be an area where (applicants) can actually affect outcomes. On all other criteria - they can't change their age, they can't vary that they did or didn't apply for refugee status by 1 November 1993 - but in English language they did have a capacity!

Despite the clearly benign experience of many candidates, there continued to be widespread objection to the test. The sitting of the first Sydney administration at the Wentworth Park Dog Track was viewed as an insult by some Chinese candidates (Sydney Morning Herald 1994b), and The Age headlined a report on test administration with 'Language Test Risks Lives, Say Chinese Students' (The Age 1994). By October 1994 a number of Class 816 applicants were considering class action against DIEA, to force a cancellation of the test based on the UN Refugee Treaties and the Privacy Act (United Chinese Times 1994b). According to a lawyer acting for the Chinese Students Union for 816, this decision was taken out of fear of a possible monitoring of tests by Chinese spies (The Australian 1994) - a concern defused once DIEA offered the option of private sittings (The Tide 1994b). Other major concerns persisted. Many 816 candidates had been exempted from qualifications assessment, given their completion of accredited Australian courses prior to application. Despite this, by September 1995 only 61 per cent of the 5,030 cases referred for specialist assessment to NOOSR had received positive outcomes - an issue of passionate concern to PRC applicants possessing two year teachers college or radio/television college degrees, who were initially ruled ineligible by NOOSR. Further class action was considered, this time against NOOSR on the presumption it was in league with DIEA to minimise the number of those passing (United Chinese Times 1995a). Specific concerns related to a perceived downgrading of Chinese qualifications in the assessment process, unwillingness to take

account of individual circumstances and potential, and differential assessment of identical qualifications (Chinese Post 1995). A further issue was the degree to which trades-qualified candidates could meet DIR requirements - with dramatically lower rates of qualifications recognition being achieved[4].

An intense lobbying of prominent Chinese community members commenced, with student offers including support for the federal leader of the Opposition, John Howard (should he win the election and back the PRC students' cause), and support for the re-election bid of Henry Tseng, Deputy Lord Mayor of Sydney (United Chinese Times 1995b). By mid-1995, as news of negative qualifications and *step* results circulated, agitation within the Chinese community rose again, together with a growing demand for all Class 816 candidates to be accepted. A small group applied for asylum on the grounds of homosexuality - only to find the burden of proof problematic (United Chinese Times 1995c). Other canvassed options included PRC students raising money to build a Sydney Olympics Stadium to prove their civic worth, or taking to the streets with ongoing demonstrations and hunger strikes (Independence Daily 1995). Such cumulative action secured support. By July 1995 the Chinese Association of Victoria - an umbrella organisation of 22 Chinese groups - passed a resolution to endorse the demand of all rejected Class 816 applicants to stay in Australia (United Chinese Times 1995d), a position also adopted by the Chinese Nationals Association (Chinese Herald 1995a).

In response to such pressures, DIEA spokespeople steadily reiterated the fairness of the test, its lack of set quotas, and the Department's determination that outcomes would be binding. People attempting the test a second time could do so in May 1995, with a further administration deferred to August to cater to candidates previously unavailable. Additional tests were scheduled for September and October 1995, as well as early 1996, to cater for emerging cases. According to Candlin, the rationale for these further administrations was the Department's desire for candidates to have an entitlement of two attempts and an appeal, should they have failed by only one module in the second attempt. DIEA was also concerned to allow for factors like illness, late applications, and successful appeals following the initial non-recognition of academic qualifications.

The results of successive *step* tests in fact authenticate DIEA claims of fairness. 78 per cent of candidates passed all three sections in the first round of testing, followed by 58 per cent in January 1995. By September 1995, according to DIEA, 9,758 candidates had sat for the test, with 88 per cent passing - a degree of success immediately leading to an upsurge in demand in China for student visas (Chinese Herald 1995b), with those failing perceived to have done so primarily due to lack of preparation. Of the total *step* candidates, around 84 per cent were former PRC students (many Sri Lankans having received exemptions, given the medium of their education had been English). When the 'November 1' telephone line closed in late 1995, it had received around 150,000 calls since its establishment in January 1994. By March 1996 8,733 Class 816 principal applicants (amounting to 11,361 persons) had gained permanent residence in Australia, together with 2,184 Class 818 applicants. 40,733 persons in all had been accepted from Class 815, 816 and 818 candidates. Of these 37,017 were from the People's Republic of China, representing 91% of approvals. Around 1,004 Class 816 applications remained to be finalised, with the assessment of qualifications proving to be a far more formidable barrier. While those missing out would revert to the process of determination of refugee status, the implementation of *step* (backed by qualifications assessment) had permitted 7,323 cases to be removed from the backlog within a two year period - almost wholly achieved on a user-pays basis.

A DIEA spokesperson summed up the value of the process in the following way:

> ..(T)he thing I can see is that this group have now got rid of their uncertainty. We will now be able to go ahead and consider claims for asylum much more quickly. There's no doubt in our minds that these people will make a substantial contribution to Australia. You just have to look at the size... of the Chinese export market. These people no doubt... will provide the bridge.

CONCLUSION

The case of the *step* test offers a dramatic illustration of the increasing use of language testing by Australian authorities to achieve political purposes. While the ultimate merits of the strategy and outcomes will remain a matter of debate, *step* (in alliance with qualifications assessment

bodies) had a capacity to deliver the Australian government a solution which was timely, cheap, and administratively simple. Three further benefits were inherent in the process. The federal government was able to reimpose control over a politically volatile situation; the Australian legal system was cleared of an unmanageably large backlog of refugee applications; and young PRC, Pakistani, Sri Lankan and other asylum seekers were transformed into a relatively educated and acculturated skilled migrant intake - commencing their legal permanent residence supported by a functional basis in English.

Macro-political issues such as those outlined above clearly have a profound potential impact on test design, administration and outcomes. I believe they warrant detailed consideration by applied linguists. Indeed, solutions such as *step* may prove irresistible under similar circumstances in the future.

NOTES

1 The dictation test was incorporated in the Commonwealth Immigration Act of 1901 and "required any intending immigrant to pass a written test, originally in 'any European langauge'. However, the language did not need to be one understood by the immigrant. The object of the test was entirely to facilitate exclusion rather than to ascertain whether immigrants were literate." (Jupp 1991:46).

2 Following the events at Tiananmen Square in 1989, Australia's then Prime Minister (Robert Hawke) gave an undertaking that no Chinese students then resident in Australia would be returned to China against their will.

3 It is important to recognise in relation to the figures provided within this chapter that there has been a degree of slippage between Class 816 and Class 818 figures over time, given transfer by candidates between categories.

4 Qualifications related to the metal, electrical and some building trades were assessed by the Department of Industrial Relations. Around 37% of these were judged as equivalent to Australian trade qualifications - a rate comparable to DIR outcomes for skill migration assessment, but low compared to NOOSR qualifications assessments (positive in around 60% of cases). Other trades (eg related to the

catering and service industries) were assessed by state and territory training authorities, given access to registration for these occupations was governed by state legislation.

REFERENCES

The Age (1994) Language test risks lives, say Chinese students. K Middleton and J Painter. 2 November 1994. Melbourne.

The Australian (1994) Chinese in court action over visas. 4 November. Sydney.

Birrell, R (1993) The politics of the refugee issue. *People And Place* Vol 1 No 4. Australian Forum for Population Studies, Melbourne, Monash University.

Birrell, R (1994a) Implications of the November 1993 and January 1994 amnesties: how many will be affected and at what cost? *People And Place* Vol 2 No 1, Australian Forum for Population Studies, Melbourne, Monash University.

Birrell, R (1994b) The outcome of the 1 November 1993 decisions. *People And Place*, Vol 2 No 3, Australian Forum for Population Studies, Melbourne, Monash University.

Birrell, R (1995) The 1995-96 migration program. *People And Place*, Vol 3 No 2, Australian Forum for Population Studies, Melbourne, Monash University.

Chinese Herald (1994) Letter to the editor, 26 September.

Chinese Herald (1995a) Article on growing support for residential status to be offered to all class 816 students, 15 September.

Chinese Herald (1995b) Article on upsurge in demand for visas following the success of many class 816 Candidates, 9 April.

Chinese Post (1995) Article on perceived malpractice in Chinese qualifications assessment, 15 June.

Commonwealth Of Australia Gazette (1993) No S 33, 1 February, Canberra

Davenport, M (1990) Here today, here tomorrow. *The Independent Monthly*, July 9-10.

DFAT (1994) *Country profile: China*. Refugees, Immigration andAsylum Section, Canberra, June.

DIEA (1988) Australia and immigration, 1788 to 1988. AGPS, Canberra.

DIEA (1992) Changes to entry arrangements for students from the People's Republic Of China (PRC), Media Release, Senator Michael Tate, 28 October.

DIEA (1994) Application form for main applicant, Form 1816, 1818, Canberra.

DIEA (1995a) 1 November 1993 Decisions, Fact Sheet 20, Revised 23 March 1995, Information and Publishing Section, Canberra.

DIEA (1995b) Seeking asylum within Australia, Fact Sheet 19, Revised 9 February 1995, Information and Publishing Section, Canberra.

DIEA (1996) Unpublished statistics on outcomes of all applications from Class 815, 816 and 818 applicants, by country of origin, March 8 1996, November 1 Taskforce, DIEA, Canberra.

Hansard (1994) Senate Proceedings. 11 October, Questions No 1614-1617, Canberra.

Hawthorne, L. (1994) The politicisation of English: the evolution of language testing. *People And Place*, Vol 2, No 2, Melbourne, Monash University.

Hawthorne, L. (1995) The politicisation of English part two: the access: test and the skilled migration program. *People And Place*, Vol 3, No 1.

Hawthorne, L. & Toth, J. (forthcoming) Immigrant nurses in Australia. AGPS, Canberra.

Independence Daily (1995) Article on alternative strategies for gaining permanent residence, 3 July.

Jupp, J. (1991) *Australian retrospectives: Immigration.* Sydney, Sydney University Press

McNamara, T. (1990) Assessing the second language proficiency of health professionals. Melbourne, University of Melbourne, Unpublished PhD Thesis.

NCELTR (1994) Information for candidates: *step* Special Test of English Proficiency, Sydney, School of English and Linguistics, Macquarie University.

Nesdale, D, Simkin, K, Sang, D, Burke, B & Fraser, S (1994) International students and immigration, AGPS, Canberra.

New Migrant (1994) Article on *step* test. September Edition.

Rubin, K. (1995) Homeward bound: China's returnees face exciting prospects - and tough challenges. *Far East Asian Economic Review* May 18 1995, Hong Kong.

Shu, J & Hawthorne, L. (1996) Asian student migration to Australia. *International Migration Quarterly Review* First Quarter 1996, Switzerland, International Organisation For Migration.

Sydney Morning Herald (1994a) Immigration English test editorial. 3 November, Sydney.

Sydney Morning Herald (1994b) Chinese refugees claim test an insult. 2 November.

The Tide (1994a) English test may be cancelled. strive to make this happen. 15 September.

The Tide (1994b) Article on termination of class action by class 816 candidates, 24 November.

United Chinese Times (1994a) Article on dangers facing Chinese returnees. 16 September.

United Chinese Times (1994b) Article on the possibility of class action forcing a cancellation of the test. 21 October, Sydney .

United Chinese Times (1995a) Article on visit of Chinese delegation to NOOSR. 2 June, Sydney.

United Chinese Times (1995b) Article on an emergency appeal from PRC students to Henry Tseng, Deputy Lord Mayor of Sydney, 7 September, Sydney.

United Chinese Times (1995c) Article on application to chance grounds on basis of homosexuality, 6 July, Sydney.

United Chinese Times (1995d) Article on unanimous support for all efforts of class 816 students to stay in Australia, 27 July.

Waldron, B (1992) The evolution of Australia's overseas student policy. Melbourne, Monash University, Unpublished M.Ed. Thesis.

York, B (1993), Admitted: 1901-1946. Immigrants and others allowed into Australia between 1901 and 1946. *Studies in Australian ethnic history*,. No 2, Centre For Immigration and Multicultural Studies, Australian National University.

DEVELOPING LANGUAGE TESTS FOR SPECIFIC POPULATIONS

Rosemary Baker
The University of Queensland

ABSTRACT

This article examines the decisions underpinning language test design in circumstances where existing tests are of questionable validity for particular subpopulations, or where no suitable test exists. Three strands of test development work are described, all concerned with the assessment of language abilities in bilinguals and speakers of English as a second language in Australia. The specific contexts discussed are: (1) the assessment of residual language-processing abilities following stroke; (2) the assessment of communicative ability following stroke; and (3) the assessment of first and second language abilities in Alzheimer's dementia. For all of these purposes, tests need to be suitable for people of a range of ethnolinguistic backgrounds and degrees of bilingualism. Each context, however, raises its own set of considerations, and has therefore necessitated a different approach. In the first case, the starting point for the development of appropriate test content has been the investigation of a published aphasia test already available in many of the required languages. In the second case, an analysis of the communicative needs of a sample of stroke patients of non-English-speaking background was conducted, and content for a new measure selected on the basis of the results. In the third case, content selection has been informed primarily by the findings of empirical research on language decline in dementia.

INTRODUCTION

Language assessment in neurological impairment

Assessment of the language abilities of people with neurological impairment serves several purposes. First, the results can inform the diagnosis of neurological disorders. As Simmons (1985) explains:

> In many ways communication is a mirror of neurologic function. Careful observation of speech, language, memory, and nonverbal communication features can signal the presence of a neurologic

> disorder, suggest etiologic factors, provide clues to the extent and location of lesions, and reflect a change in neurologic status over time. Communication assessment is therefore an important component in the accurate diagnosis and treatment of neuropathologies.
>
> (Simmons 1985:4)

Second, such assessments form the basis for the speech pathologist's management of affected individuals. As Shipley and McAfee (1992:ix) note, assessment information is used to make diagnoses, identify the need for treatment and for referral to other professionals, and to determine the focus, frequency, length and structure of treatment. Wirz (1993:11-12) distinguishes two types of assessment preceding remedial intervention: (1) descriptive assessment, which describes the individual's communicative behaviour in terms of its departure from known norms, and (2) prescriptive assessment, which focuses on the problems which are remediable and identifies possible starting points for therapy. She further distinguishes between baseline measures taken before therapy and outcome measures taken later, noting that these are sometimes combined in studies of the efficacy of intervention.

Third, the identification and analysis of the linguistic correlates of various neuropathologies have an important research function. Beginning from the pioneering work by Dax, Broca and Wernicke in the 1800s, on the localisation of language in the brain, such investigations continue to contribute to understanding of the cerebral representation of language, of aspects of language processing, and of the disorders which result in its disturbance.

There exists a wide range of procedures for the assessment of different aspects of neurogenic communication disorder, including inventories and questionnaires, analyses of speech samples, instrumental measures, informal observations and formal tests. Two points regarding the formal, published tests are of importance in the present context. First, they have been developed for particular populations (predominantly in the USA), and their validity for populations other than these is known only insofar as additional normative data have been gathered (see, for example the Australian data on the Boston Naming Test provided by Worrall, Yiu, Hickson and Barnett (1995)). Second, most assume that

the patient or subject is unilingual, or at least that testing in one language will provide an adequate indication of the person's residual linguistic and communicative abilities.

This article is concerned with the development of language tests for populations for whom this is not necessarily the case, namely, bilinguals and speakers of English as a second language resident in Australia. Approaches which would take account of the characteristics of some of these populations were proposed in a previous article (Baker 1993). The work described here represents the next step towards the design of appropriate procedures in the areas of aphasia assessment (specifically, the assessment of residual linguistic processing and communicative abilities after stroke), and the assessment of language disorder in Alzheimer's dementia. In each of the three testing contexts discussed below the starting point for the development of a suitable approach has been a consideration of existing published tests and their potential applicability to the populations of interest.

Characteristics of the target populations

The 1991 census showed that in Australia, approximately 15% of people aged 5 and over speak a language other than English at home (Australian Bureau of Statistics 1993). The twelve most commonly spoken languages specified were Italian, Greek, Chinese languages, Arabic, German, Vietnamese, Spanish, Polish, Macedonian, Croatian, Maltese and Dutch. Almost 80% of speakers of languages other than English reported themselves also to be able to speak English either well or very well.

In the elderly sector of the population, the proportion of speakers of languages other than English is higher, with more than 17% of people aged 65 and over having been born in non-English-speaking countries (Australian Bureau of Statistics 1993). Of these elderly people, over 40% reported themselves as being either unable to speak English or unable to speak it well.

The National Heart Foundation of Australia (1995) reports that of the estimated average of 100 new strokes occurring every day in Australia, approximately 60% are in people aged 65 and over. Henderson and Jorm (1986) report that the prevalence rates of moderate and severe dementia in Australia rise from 1.6% in those aged 65-69 to 28.7% in those aged over 85, approximately doubling every five years in between.

In approaching the development of language tests, then, it is necessary to take account of the age-related nature of the conditions underlying the language disorders to be assessed, as well as the considerable variation in degrees of bilingualism found within the target populations.

ASSESSMENT OF LANGUAGE-PROCESSING ABILITIES FOLLOWING STROKE

Aphasia in bilinguals and polyglots

Reported cases of aphasia in speakers of more than one language show that patterns of impairment and recovery in a stroke patient's different languages are unpredictable. As Paradis (1995a:212) explains, the language first or best recovered has not proved to be consistent either with the 'rule of Ribot' (that it is the language first acquired), or with the 'rule of Pitres' (that it is the language most familiar at the time of cerebral insult), or indeed with subsequent suggestions that it is the one with the strongest affective ties or the most useful one post-stroke. Paradis (1995b:7) further points to the possible contribution of explicit metalinguistic knowledge gained during formal instruction to the apparent preferential recovery of a second or foreign language in which the person was not even fluent before insult. His explanation for this is that explicit knowledge of this kind would still be available to the person post-onset, even though access to implicit linguistic knowledge (including that of the first language) might be impaired.

As regards the different patterns of recovery observed over the last hundred years, Paradis (1993:278-9) describes six basic types: parallel (all languages recovered equally and at the same time), differential (one language recovered better than another), successive (one language after another), selective (not all languages recovered), antagonistic (the language initially recovered is replaced by another), and mixed (the person mixes elements of different languages). Given the variety and unpredictability of possible outcomes in terms of the accessibility of people's different languages following stroke, it is important that appropriate procedures be available for assessing patients' performance in all of their languages. Indeed, lack of adequate assessment in languages other than that of the immediate hospital environment carries a risk of misdiagnosis.

Aphasia test batteries

Kerr (1993:101-6) notes that the content of formal test batteries over time has reflected prevailing theories of aphasia. Tests such as the Boston Diagnostic Aphasia Examination (BDAE) (Goodglass and Kaplan 1983) and the Western Aphasia Battery (WAB) (Kertesz 1982) emphasise the categorization of patients according to clinical aphasia syndromes. Meth, Obler and Walsh (1993:195) have pointed out that although it is impossible to devise an aphasia test without making assumptions about language, authors of aphasia tests have rarely made these explicit.

Aphasia tests based on syndrome classification have been called into question in recent years for their inability to elucidate the underlying nature of the language disorder (Byng, Kay, Edmundson and Scott (1990:72). An alternative approach is evolving, in which aphasic symptoms are investigated with reference to explicit models of language processing. Such an approach is exemplified by the Psycholinguistic Assessments of Language Processing in Aphasia (PALPA) (Kay, Lesser and Coltheart 1992), which consist of a set of procedures designed to allow the investigator to specify which modules and pathways of a person's language-processing system have become impaired. Thus, for example, in testing the ability to repeat words, one would present both real words and non-words to check whether both the lexical route (phonological input lexicon to phonological output lexicon) and non-lexical route (direct acoustic-to-phonological conversion) were intact (Kay et al. 1992:11). The authors regard clinical language assessment as "an iterative procedure of hypothesis testing" (p.1).

Selection of an aphasia test for the multilingual setting

Aphasia batteries such as the BDAE and the WAB, both of which originated in the USA, have come into wide international use. The BDAE was the main comprehensive aphasia test found to be in use in hospitals in the Brisbane metropolitan area (Smith Worrall and Burtenshaw 1990), and is also reported to be the most widely-used aphasia assessment in New Zealand (Larkins 1995). However, the dangers and difficulties of attempting to translate existing formal language tests for use with populations of other ethnolinguistic backgrounds have been pointed out (Baker 1993:267; Cheng 1993:67). Furthermore, in developing

assessment procedures for use in the multilingual setting, it seems preferable to adopt an approach which is more consistent with recent thinking and understanding of language processing, and for which the psycholinguistic basis is more explicit.

An existing comprehensive aphasia test which implements, at least to some extent, an approach of this type, and which is already available in many of the languages spoken in Australia, is the Bilingual Aphasia Test (BAT) (Paradis 1989). In view of recent challenges to the theoretical status of classical aphasia syndromes, this test does not seek to categorize patients; profiles of performance on the various subtests of the BAT can, however, be interpreted in these terms if the clinician so chooses (Paradis 1987:211). The description of its theoretical foundations (Paradis 1987:32-3) also suggests that it is intended to be suitable for people whose educational and language proficiency levels may be quite modest. It therefore seemed to offer a useful starting point for the development of appropriate procedures for assessing residual language-processing abilities following stroke in speakers of these languages in the Australian setting.

Investigation of applicability

Since the content of an aphasia test should present little or no difficulty to people with no neurological impairment, the first step in assessing the potential applicability of the BAT to people of non-English-speaking background in the Australian setting is to administer it to normal members of these populations. An investigation of this type carried out with normal elderly members of five of the most common language backgrounds (Baker, 1995a) is summarised below.

A total of 93 people aged between 63 and 85 (mean age 72) participated in this study. All were tested using the Bilingual Aphasia Test in their first language (Cantonese, Dutch, Greek, Polish or Vietnamese). The English version of the BAT was also administered to 69 of the subjects; all but one of those who did not take part in the English testing were unable or unwilling to do so because of low English proficiency. Upon examination of the scores from each of the 32 subtests of the BAT, for all languages tested, three categories of task could be discerned, with approximately equal numbers of tasks in each category: (1) those on

which a high proportion (over 75%) of subjects scored highly both in their first language and in English, (2) those on which scores were generally high in the first language but not in English, and (3) those on which fewer than 75% of people scored highly in either language.

The tasks which were found to be unproblematic both in people's first languages and in English were: naming and recognising the names of common objects, following simple commands, constructing short semantically-acceptable sentences from given words, answering factual questions on a short listening comprehension passage, reading words and sentences aloud, copying words, word reading comprehension using pictures, and reciting series such as days of the week.

The tasks which were performed well by over 75% of people in the first language, but caused more difficulty in English, were: following three-stage commands, constructing short grammatically-correct sentences from given words, judging the semantic acceptability of sentences, auditory discrimination of single words using pictures, recognition of semantic categories in verbally-presented multiple-choice items, repetition of words and sentences, production of semantic opposites, judgement of the semantic acceptability of sentences, factual comprehension questions on a short reading passage, comprehension of reversible noun phrases, and word dictation. From the results on these subtests in people's first languages, it appeared that the tasks themselves were manageable, but that they made greater demands when administered in English. An obvious contributory factor to this is the level of English proficiency required. However, the possible influence of some additional variables is considered below.

The third category of subtests were those on which performance was generally least good, both in the first language and in English, suggesting that problems were caused by the nature of the tasks themselves. These included a very long oral syntactic comprehension section involving selection of pictures showing the correct thematic roles for affirmative and negative sentences of a variety of structures, the identification of synonyms and antonyms in verbally-presented multiple-choice items, grammaticality judgement, lexical decision, sentence dictation, sentence reading comprehension (using sentences of the same types as in the oral syntactic comprehension section), mental arithmetic, and items based on derivational morphology.

Possible relationships between BAT subtest scores and other variables were explored using correlational analysis, with the significance level set at .001 in each case. Not surprisingly, level of English proficiency, as rated on a 0-5 scale by the English-speaking examiner on the basis of the subject's spontaneous speech, showed a significant positive correlation with performance on all but three of the BAT subtests in English (pointing to named objects, reciting series and copying). No significant associations were found between either age or hearing (as measured by pure-tone audiometry) and performance on any of the BAT subtests in people's first languages. In English, however, decreasing performance on three subtests was associated with advancing age: following three-stage commands, recognising synonyms in orally-presented multiple-choice items, and constructing semantically acceptable sentences from words presented orally. Further, decreasing performance on the auditory discrimination and word repetition tasks in English was associated with increasing levels of hearing impairment.

Paradis (1987:32) acknowledges that some of the BAT subtests are sensitive to years of education. In this study, level of education was found to be significantly positively correlated with scores on 12 subtests: in the case of sentence construction (semantic acceptability), word dictation, mental arithmetic and generating words with given initial sounds, this was true both of first language and English performance; for grammaticality judgement and synonym, antonym and semantic category recognition it applied to first languages only; and for sentence reading, syntactic comprehension of affirmative sentences and morphology tasks, it applied only to English.

These results thus indicate which parts of the BAT would be appropriate in their current form for assessing aphasia in stroke patients of similar ages and backgrounds to the subjects in this study. They also point to some potential sources of invalidity which will need to be avoided by omission or modification of certain subtests, and by recognition of extraneous factors that might affect scores. The results suggest that when testing is carried out in the second language, performance might be influenced to a greater extent than in the first language by age (particularly on items which place heavy demands on memory) and age-related hearing impairment (particularly on items where there is no supporting context).

It is possible that more of the content of the BAT would be appropriate for younger members of these communities, or members with higher educational levels than tend to be found among the current generation most at risk of cerebro-vascular disease. However, because of the marked cohort differences between the generations, particularly in terms of degree of bilingualism and educational level, it was necessary to select for this study a subject sample which was relatively homogeneous in terms of age. The applicability of the BAT with younger bilinguals therefore remains to be checked empirically.

ASSESSMENT OF COMMUNICATIVE ABILITY FOLLOWING STROKE

Tests of communicative ability

In speech pathology, communicative ability in everyday life is usually referred to as 'functional communication' or 'communicative function'. Published tests of everyday communicative ability include the Functional Communication Profile (FCP) (Sarno 1969), the Communicative Abilities in Daily Living (CADL) (Holland 1980), the Communicative Effectiveness Index (CETI) (Lomas, Pickard, Bester, Elbard, Finlayson and Zoghaib 1989), and the Amsterdam-Nijmegen Test of Everyday Language Abilities (ANELT) (Blomert, Kean, Koster and Schokker 1994). An indication of the content of each of these is given in Table 1 below. Of the four measures shown, only the CETI was based on prior investigation of the communicative needs of its target population: groups of aphasic patients and their spouses were asked to generate a list of communicative situations that they considered important in their everyday lives, and the test content was drawn from the most generalizable of these. The rationale for the content of the FCP, on the other hand, was stated simply as being that the communication behaviours included were "considered common functions of every day urban life" (Sarno 1969:2). Similarly, the content of the CADL was assumed to represent "living features common to 'most' Americans" (Holland 1980:2). Blomert et al. (1994:4) report that the situations chosen for the ANELT were judged by 60 normal adults as being "highly imaginable and recognizable, independent of biographical background, and

representative of daily life situations". No data on the plausibility of these situations for persons with aphasia are provided, however.

Suitability of existing measures

Two limitations on the applicability of existing measures of everyday communicative ability to the Australian multilingual setting can be noted from these examples. First, the content of the ANELT and CADL focuses on communication outside the home, in settings such as the doctor's office and shops. Indeed, the ANELT is explicitly restricted to interaction in the public domain, on the grounds that "...other domains such as family interactions are prone to contain idiosyncratic elements" (Blomert et al. 1994:4). However, a study of the communicative needs of 72 normal elderly members of five different language backgrounds in Australia indicated that the participants' main communicative purposes were in the family and social domains (Baker 1995b), thereby, for elderly populations at least, calling into question this emphasis on communication in public settings. Even in the absence of specific evidence that content might be inapplicable, however, it should not simply be assumed that test content will be valid for populations other than those for whom it was intended.

The second point concerns the practical constraints on methods of administration in the multilingual context. For languages other than English, it is rarely possible for patients to be assessed by professionals of the same language background. One therefore needs to consider the use of ratings by significant others who can report on the patient's communication (as in the CETI). Further, for the purposes of obtaining a full picture of people's residual communication abilities, a procedure is required which would provide information on the aphasic person's communication both in English and in other languages. Assessment also needs to take into account the usual domains of use of each language.

For the assessment of communicative ability following stroke, then, no single existing procedure appeared immediately suitable in both content and format. The approach taken in this case, therefore, was to determine appropriate test content on the basis of an analysis of the communicative needs of a sample of stroke patients of non-English-speaking background, to select a feasible method of administration, and to gather data on the potential ease of application and reliability of a new

Test	*Content type, with examples*
FCP (Sarno 1969)	Speech pathologist rates patient (without any set questions) on 45 items grouped under 5 headings: *Movement*, e.g. ability to indicate "yes" and "no"; indicating floor to elevator operator. *Speaking*, e.g. saying greetings; saying noun-verb combinations; speaking on the telephone. *Understanding*, e.g. awareness of emotional voice tone; understanding verbal directions; understanding rapid complex conversation. *Reading*, e.g. reading single words; reading magazines. *Other*, e.g. writing name; handling money.
CADL (Holland 1980)	Speech pathologist rates appropriacy of responses in situations in public domains, including the assessment interview itself (e.g. giving/correcting personal details); role-play of a visit to a doctor (e.g. dealing with receptionist, form-filling, recognising an implausible instruction from the doctor); car travel (e.g. reading speedometer); shopping (e.g. finding an item, using a drinks vending machine); using the telephone (e.g. finding number in directory, dialling). Interpretation of sounds, and pictures showing e.g. events, facial expressions, metaphor.
CETI (Lomas et al. 1989)	A significant other rates patient on 16 situations which require stroke survivors to be able to understand and to make themselves understood. E.g. getting somebody's attention; communicating physical problems such as aches and pains; starting a conversation with people who are not close family; being part of a conversation when it is fast and there are a number of people involved.
ANELT (Blomert et al. 1994)	Speech pathologist asks patient what to say in 10 hypothetical situations in public domains. E.g. complaining at the dry cleaner's about a scorch mark on a shirt; asking a neighbour to come and visit; arranging to have a bouquet of flowers delivered to a friend.

Table 1: Content of some published tests of functional communication

measure of this kind. This process, reported by Woodman and Baker (1995), is summarised below.

Development of an approach for the multilingual setting

In this work, particular attention was paid to the need to establish content validity on the basis of investigation of the communicative needs and patterns of bilingual or second language use of members of some of the target populations. Ten people with aphasia, ranging in age from 35 to 77 (mean age 62 years), from a variety of backgrounds, including Chinese, Dutch, El Salvadorean, Finnish, Italian, Polish and Russian, participated in this study. Three data collection procedures were designed to determine their current and pre-stroke everyday communicative needs and activities (both in terms of importance and frequency), and the language(s) they used for different purposes both currently and premorbidly.

The three methods of needs data collection were: (1) individual interviews, designed to elicit these aphasic subjects' own perceptions of the communicative purposes that were important in their daily lives; (2) questions as to whether or not they ever performed (both pre- and post-stroke) each of the items on a checklist of 57 communicative activities drawn from existing tests of functional communication and needs assessment procedures; and (3) communication diaries, completed by the subjects (with help from their significant others if necessary) for a period of 5 days, to provide a record of the interlocutors, settings, topics and modalities involved in the communicative activities which arose each day, and of any difficulties which the person experienced. These diaries were devised both as a practicable substitute for direct observation of subjects in their home environments, and as a check on the accuracy of reponses given on the checklist questions.

From the information obtained using these three procedures, a preliminary set of 19 communicative activities was selected as the basis for a new measure. The activities were chosen to be generally applicable to the subjects, both in their frequency and importance, and to sample a range of communicative situations in terms of channel (receptive, productive or both), domain, interlocutor, modality (e.g. non-verbal, speaking, writing), pragmatic function and topic. Consideration was also given to the applicability of items to people with differing levels of

severity of communication impairment, and to their potential for differentiating pre- and post-stroke communicative ability. Examples of the communicative activities included in this core set were: showing that you understand what is being said; answering questions about family; asking where items are in a shop or supermarket; listening to and understanding medication instructions.

A format modelled closely on the CETI (Lomas et al. 1989) was chosen for the new measure, to allow patients' communication to be rated by those with whom they have the most contact, and who speak the same language(s). This took the form of a 10cm visual analogue scale, with the words "Same as before stroke" at one end, and "Cannot" at the other (either in English or in other languages), upon which the significant other would place a mark to indicate the patient's current versus premorbid ability on each item, in whichever language would normally be used for that activity.

To investigate the feasibility of this new approach, and to gather preliminary data about its potential reliability, the same 10 (stable) aphasic subjects were rated using the 19 items by their significant others on two separate occasions approximately two weeks apart. The ratings were converted to numeric scores, and compared for the two administrations. The pairs of ratings showed a promising degree of stability ($r=0.74$, $p<.0005$), given that the item set and instructions to raters were in their preliminary form. Further analysis of the results showed that four items in particular were not reliably rated, and that two raters showed particular inconsistency across occasions; this information will be used in the selection and modification of the questions and instructions for the measure. In general, however, the results support the continued development and refinement of this approach.

ASSESSMENT OF LANGUAGE ABILITIES IN ALZHEIMER'S DEMENTIA

Language assessment in dementia

The final area of work discussed here concerns the development of materials to investigate first and second language abilities in people with dementia. In addition to the theoretical interest of describing and

elucidating changes in language performance in age-related disease, such information has important practical implications for the provision of appropriate aged care programs and health services for the ethnic aged.

Kertesz (1994) provides the following rationale for language assessment in dementia:

> Language deterioration is an important sign in early Alzheimer disease, and it is a major symptom needing assessment in the majority of patients reaching the disease's middle stages. The monitoring of language change becomes an important tool in following the course of dementia and the effect of possible therapy, although important dissociations can be observed between various elements of cognition, such as language, memory, and visuospatial functions. Language changes are important to distinguish between etiologically distinct varieties of degenerative disease (such as Alzheimer and Pick disease) or vascular dementia. In the final stages of the disease, when memory and nonverbal cognition cannot be tested because of floor effects, elicited language performance may be the last standardizable, psychometric instrument available to quantitate change.
>
> (Kertesz 1994:134)

Research on the nature of language impairment in Alzheimer's dementia is reviewed by Miller and Morris (1993) and Melvold, Au, Obler and Albert (1994). In brief, the research findings to date suggest that the areas of earliest and most marked decline are those dependent on lexical semantic and conceptual knowledge, as evidenced by impaired ability to name objects or pictures and comprehension difficulties. Phonology and syntax are thought to remain relatively stable, at least until the late stages of dementia. Such findings are of importance to the present discussion for the information they provide about the aspects of linguistic performance that might (1) differentiate persons with dementia from normal elderly control subjects, and (2) differentiate dementia patients in terms of their stage of severity.

The language assessment of persons with dementia does not, in general, form a routine part of the speech pathologist's work, though differential diagnosis of aphasia and dementia is sometimes required. There is not the abundance of published materials for the assessment of language in dementia that one finds for the assessment of aphasia. Indeed, test materials used in the reported research on dementia are sometimes

drawn from aphasia tests. There is, however, one test battery designed specifically to assess language impairment in dementia: the Arizona Battery for Communication Disorders of Dementia (ABCD) (Bayles and Tomoeda 1991).

The content of the ABCD was selected primarily on the basis of research findings concerning language decline in dementia. Bayles and Tomoeda (1991:39) note that all but four of the thirteen task types had previously been used in longitudinal studies of people with Alzheimer's dementia. In addition to subtests evaluating mental status and drawing, the ABCD contains the following language-based tasks: immediate and delayed story recall, word recall (free, cued, and by recognition), object description, generative naming by semantic category, naming objects in black-and-white line drawings, word definition, following commands, comparative questions (e.g. *Is ice colder than steam?*), repetition of semantically anomalous phrases and sentences, and picture-based reading comprehension of words and sentences. The authors claim (p.39) that the ABCD can differentiate mild from moderate Alzheimer's disease, and that it is therefore useful for documenting disease effects over time. They suggest also that the word recall and delayed story recall subtests can be used as screening tools to identify people who may be in the early stages of dementia.

Selection and adaptation of tasks for the multilingual setting

Case studies of language abilities in bilingual or multilingual speakers with Alzheimer's dementia are reported by Hyltenstam and Stroud (1989) and de Santi, Obler, Sabo-Abramson and Goldberger (1990). The elicitation procedures used in the study by Hyltenstam and Stroud were conversation on personal topics and on the immediate environment, the picture description task from the Boston Diagnostic Aphasia Examination (Goodglass and Kaplan 1983), action picture naming from the Action Naming Test (Obler and Albert 1986), and an automatic speech task similar to that in the BDAE (counting to 20, and reciting the months of the year and days of the week). The study by de Santi et al. (1990) also used subtests from the BDAE, including the conversation and expository speech section (which contains the picture description task), automatic speech (days, months, numbers 1 to 21 and letters of the alphabet), oral

reading and reading comprehension (multiple-choice completion items). The authors do not specify which of the repetition and writing tasks from the BDAE were administered, and indeed report no results for the repetition. The remaining procedures used in this study were the collection of spontaneous speech samples, and administration of the Boston Naming Test (Kaplan, Goodglass and Weintraub 1983), in which the patient names objects depicted in black-and-white line drawings, and the Action Naming Test (Obler and Albert 1986). In both studies, the same tasks were administered in the required languages (Finnish, German and Swedish in the first case; English and Yiddish in the second).

In the study by Sasanuma, Itoh, Watamori, Fukuzawa, Sakuma, Fukusako and Monoi (1985), on elderly people in Japan with and without dementia, the following language-based tasks were administered: immediate and delayed story recall, sentence repetition (from a Japanese aphasia test), following commands, picture naming, generative naming by semantic category (animals) and by initial syllable, picture description, oral reading and reading comprehension of single words in Kana and Kanji, and sentence dictation. It was found that the performance of the subjects with dementia (a combined group of cases of different aetiologies) on all tasks was significantly inferior to that of the normal controls, but that the degree of deterioration varied. The language tasks which seemed to be the most sensitive to dementing illness were immediate and delayed story recall and generative naming (by both semantic and phonological indices, though the phonological index appeared to be more sensitive to normal ageing). The language tasks which showed least deterioration were sentence repetition, oral reading and reading comprehension (single words).

Lee, Woo and Chen (1991) report on the use of a Chinese "anomalous sentences test", involving the repetition of semantically anomalous sentences of increasing length. This test, which was adapted from an English version, contains sentences such as "Weren't the slow slippers robbed by the big fly?" The authors conclude from their results that this task would be sufficiently sensitive to be used as a screening tool for differentiating dementia from depression in the Hong Kong Chinese elderly.

A study in which test tasks were prepared for use with a person of Japanese background with suspected dementia in the Australian setting

is reported in Baker (1996). The process of task selection and adaptation was informed in this case by findings of the type presented above, but with recognition that for migrants, the applicability of tasks used in other settings, including the country of origin, will be influenced in unpredictable ways by factors such as their degree of acculturation to the country of residence.

Consideration was given to the task-types that might be appropriate, given the (limited) available prior knowledge about the linguistic and cultural background of the subject, and her personal history, e.g. the era during which her schooling would have taken place, and therefore the type of writing that she would have learnt. At the same time, an attempt was made to include tasks reported in the literature to be particularly sensitive to the presence and severity of dementia (e.g. naming objects from drawings, story recall and generative naming by semantic category) as well as some which would tap abilities typically spared until the late stages (e.g. reciting the days of the week).

It was suggested in a previous article that the Bilingual Aphasia Test might be of use in the assessment of language impairment in dementia (Baker 1993). However, in view of its length, and the difficulty of some of its content for the ethnic aged in Australia, even in their first languages and in the absence of neurological impairment, one would need to make a careful selection of items for subjects whose attention span and concentration were limited by dementing illness. In the study reported in Baker (1996), some of the task content was indeed drawn from the Japanese version of the BAT (Paradis and Hagiwara 1989), but the task-types themselves were modelled more closely on the ABCD (Bayles and Tomoeda 1991).

CONCLUSION

The work described in this article illustrates some ways in which decisions involved in the design of procedures for assessing language impairment have sought to take account of linguistic and cultural diversity in the target populations, while also being appropriate within their clinical context. Specifically, it exemplifies (1) an approach to test development informed by prior investigation of an existing procedure; (2) test content selection based on an analysis of communicative needs; and (3) content

selection based on the findings of empirical research reported in the literature.

The investigation of the Bilingual Aphasia Test has indicated which parts of the test are appropriate in content and difficulty for elderly members of some Australian migrant populations, and which parts would need to be omitted or modified to take account of the differences between these subpopulations and the populations for whom the test was originally designed. The performance of people from other language backgrounds and from younger age groups remains to be investigated. In view of the length of the current BAT, development of appropriate screening versions would also be desirable.

Investigation to date of the proposed method for assessing everyday communicative ability in people of different language backgrounds suggests that such an approach would be feasible, and would allow the speech pathologist to obtain valuable information which would otherwise not be available. The gathering of communicative needs data from a wider variety of stroke patients of non-English-speaking background, and further trialling of a refined version of the measure, are now indicated.

Consideration of the selection and development of test materials for use in investigating first and second language abilities in dementia highlights the need to take account of the linguistic, cultural, educational and clinical characteristics of the target populations. It also points strongly to the need for appropriate normative data to aid in the interpretation of individual performance. Longitudinal studies of language decline in individuals are also required, so that the relative vulnerability of first and second languages at each stage of dementing illness can be documented. Further work is therefore envisaged in these areas.

ACKNOWLEDGEMENTS

The contribution of the following people to subject recruitment and data collection is gratefully acknowledged: Lorena Butcher, Hiroko Cockerill, Dace Dambergs, Margaret Hess, Hrissagvi Mirkovic, Chuong Nguyen, Kim Pham, Frank Rutyna, Greta Sloot, Lien Tran and Anna Woodman. Helpful discussion of the design of the needs analysis with J. Charles Alderson of Lancaster University was made possible by a British Council Academic Interlinks Scheme travel fellowship.

REFERENCES

Australian Bureau of Statistics (1993) *1991 Census of population and housing*. Canberra: ABS.

Baker, R. (1993) The assessment of language impairment in elderly bilinguals and second language speakers in Australia. *Language Testing* 10, 3:255-276.

Baker, R. (1995a) Applicability of the Bilingual Aphasia Test in multilingual Australia. Paper presented at the International Symposium on Communication Disorders in Bilingual Populations, Haifa, Israel, August 1995.

Baker, R. (1995b) Communicative needs and bilingualism in elderly Australians of six ethnic backgrounds. *Australian Journal on Ageing* 14, 2:81-88.

Baker, R. (1996) Language testing and the assessment of dementia in second language settings: a case study. *Language Testing* 13, 1:3-22.

Bayles, K.A. and C.K. Tomoeda (1991) *Arizona Battery for Communication Disorders of Dementia*. Tucson, AZ, Canyonlands Publishing.

Blomert, L., M-L. Kean, Ch. Koster, and J. Schokker (1994) Amsterdam-Nijmegen Everyday Language Test: construction, reliability and validity. *Aphasiology* 8:381- 407.

Byng, S., J. Kay, A. Edmundson and C. Scott (1990) Aphasia tests reconsidered. *Aphasiology* 4: 67-91

Cheng, L-R.L. (1993) Asian-American cultures. In Battle, D.E. (ed.) *Communication disorders in multicultural populations*. Boston, Andover Medical Publishers.

Goodglass, H. and E. Kaplan (1983) *Boston Diagnostic Aphasia Examination*. Philadelphia, Lea & Febiger.

Henderson, A.S. and A.F. Jorm (1986) *The problem of dementia in Australia*. Canberra, Australian Government Publishing Service.

Holland, A.L. (1980) *Communicative Abilities in Daily Living: a test of functional communication for aphasic adults*. Baltimore, University Park Press.

Hyltenstam, K. and C. Stroud (1989) Bilingualism in Alzheimer's dementia: two case studies. In Hyltenstam, K. and L.K. Obler (eds.) *Bilingualism across the lifespan*. Cambridge, Cambridge University Press.

Kaplan, E., H. Goodglass and S. Weintraub (1983) *Boston Naming Test*. Philadelphia, Lea and Febiger.

Kay, J., R. Lesser and M. Coltheart (1992) *Psycholinguistic Assessments of Language Processing in Aphasia (PALPA)*. Hove, East Sussex, Lawrence Erlbaum.

Kerr, J. (1993) Assessment of acquired language problems. In Beech, J. R. and L. Harding (eds) *Assessment in speech and language therapy*. London, Routledge.

Kertesz, A. (1982) *The Western Aphasia Battery*. New York, Grune and Stratton.

Kertesz, A. (1994) Language deterioration in dementia. In Emery, V.O.B. and T.E. Oxman (eds) *Dementia: presentation, differential diagnosis and nosology*. Baltimore, Johns Hopkins University Press.

Larkins, B. (1995) Aphasia treatment in New Zealand. Results of a nationwide survey. Paper presented at the Aphasiology Symposium of Australasia, Brisbane, September 1995.

Lee, H.B., J. Woo and H-C. Chen (1991) Preliminary development of the Chinese anomalous sentences test: a screening tool for the Hong Kong elderly. *Clinical Gerontologist* 11, 2:13-26.

Lomas, J., L. Pickard, S. Bester, H. Elbard, A. Finlayson and C. Zoghaib (1989) The Communicative Effectiveness Index: development and psychometric evaluation of a functional communication measure for adult aphasia. *Journal of Speech and Hearing Disorders* 54:113-24.

Melvold, J.L., R. Au, L.K. Obler and M.L. Albert (1994) Language during aging and dementia. In Albert, M.L. and J.E. Knoefel (eds) *Clinical neurology of aging* (2nd edition). New York, Oxford University Press.

Meth, M., L. Obler and P. Walsh 1993 Linguistic principles underlying two aphasia tests. In Paradis, M. (ed.) *Foundations of aphasia rehabilitation*. Oxford, Pergamon Press.

Miller, E. and R. Morris (1993) *The psychology of dementia*. Chichester, John Wiley & Sons.

National Heart Foundation of Australia (1995) *Heart and stroke facts report*. Canberra, National Heart Foundation of Australia.

Obler, L.K. and M. Albert (1986) *Action Naming Test*. Boston, VA, Medical Center.

Paradis, M. (1987) *The assessment of bilingual aphasia*. Hillsdale, N.J., Lawrence Erlbaum.

Paradis, M. (1989) *Bilingual Aphasia Test*. Hillsdale, N.J., Lawrence Erlbaum.

Paradis, M. and H. Hagiwara (1989) *Bilingual Aphasia Test (Japanese version)*. Hillsdale, N.J., Lawrence Erlbaum.

Paradis, M (1993) Multilingualism and aphasia. In Blanken, G., J. Dittmann, H. Grimm, J.C. Marshall and C-W. Wallesch (eds) *Linguistic disorders and pathologies*. Berlin, Walter de Gruyter.

Paradis, M. (1995a) Epilogue: Bilingual aphasia 100 years later: consensus and controversies. In Paradis, M. (ed.) *Aspects of bilingual aphasia*. Oxford, Pergamon Press.

Paradis, M. (1995b) Introduction: The need for distinctions. In Paradis, M. (ed.) *Aspects of bilingual aphasia*. Oxford, Pergamon Press.

de Santi, S., L.K. Obler, H. Sabo-Abramson and J. Goldberger (1990) Discourse abilities and deficits in multilingual dementia. In Joanette, Y. and H.H. Brownell (eds) *Discourse ability and brain damage*. New York, Springer-Verlag.

Sarno, M.T. (1969) *The Functional Communication Profile: manual of directions*. New York, Department of Physical Medicine and Rehabilitation, New York Medical Center.

Sasanuma, S., M. Itoh, T.S. Watamori, K. Fukuzawa, N. Sakuma, Y. Fukusako and H. Monoi (1985) Linguistic and nonlinguistic abilities of the Japanese elderly and patients with dementia. In Ulatowska, H.K. (ed.) *The aging brain: communication in the elderly*. London, Taylor and Francis.

Shipley, K.G. and J.G. McAfee (1992) *Assessment in speech-language pathology: a resource manual*. San Diego, CA, Singular Publishing Group.

Simmons, N.N. (1985) The interaction between communication and neurologic disorders. In Darby, J.K. (ed.) *Speech and language evaluation in neurology*. Orlando, Grune and Stratton.

Smith Worrall, L. and E.J. Burtenshaw (1990) Frequency of use and utility of aphasia tests. *Australian Journal of Human Communication Disorders* 18, 2:53-57.

Wirz, S.L. (1993) Introduction: historical considerations in assessment. In J.R. Beech and L. Harding (eds) *Assessment in speech and language therapy*. London, Routledge.

Woodman, A. and R. Baker (1995) An approach to the assessment of functional communication in adults with aphasia in a multilingual setting. Paper presented at the National Conference of the Australian Association of Speech and Hearing, Brisbane, May 1995.

Worrall, L.E., E. M-L. Yiu, L.M.H. Hickson and H.M. Barnett (1995) Normative data for the Boston Naming Test for Australian elderly. *Aphasiology* 9:541-551

DEVELOPING RATING SCALES FOR THE ASSESSMENT OF SECOND LANGUAGE PERFORMANCE[1]

Carolyn E. Turner
McGill University, Montreal
John A. Upshur
Concordia University, Montreal

ABSTRACT

The two most common approaches to rating second language performance pose problems of reliability and validity. An alternative method utilizes rating scales that are empirically derived from samples of learner performance; these scales define boundaries between adjacent score levels rather than provide normative descriptions of ideal performances; the rating process requires making two or three binary choices about a language performance being rated. A procedure, that consists of a series of five explicit tasks, is used to construct a rating scale. The scale is designed for use with a specific population and a specific test task.

A group of primary school ESL teachers used this procedure to make two speaking tests, including elicitation tasks and rating scales, for use in their school district. The tests were administered to 255 sixth grade learners. The scales were found to be highly accurate for scoring short speech samples, and were quite efficient in time required for scale development and rater training. Scales exhibit content relevance in the instructional setting. Development of this type of scale is recommended for use in high-stakes assessment.

INTRODUCTION

In this paper, we first describe the major types of rating scales used in the assessment of second language abilities and point out some of their limitations. We then describe in detail an alternative type of rating scale and the procedure for constructing it. This procedure was followed by a group of ESL teachers in developing an oral proficiency measure. The scales that they generated are illustrated. Their accomplishments and the results of their tests demonstrate the potential utility of this type of scaling.

ARAL Series S Nº 13 (1996) 55-79

Background

The use of rating scales in second language evaluation and description has been with us for many decades. For many years, however, the dominant method of assessment has been the use of decontextualized written test items. With the arrival of the communicative era the criteria for assessment have changed. This has been manifested in a shift from discrete test items to ratings of more "natural" performances. At the same time the basis for rating performances changed from attention to formal accuracy to concern with appropriate use and effectiveness (see, e.g., Shohamy 1995). Rating scales designed to reflect this change were needed. There have been important developments, but serious scaling problems — requiring major improvement — remain.

Two general "families" of rating scales respond to the need for communicative measurement and are in common use. The first of these families includes *holistic* and *analytic* scales based upon a general theory of the development of language abilities. They are generic and are applied to performances on different tasks by different groups for different evaluation purposes. The second family incorporates the *primary trait* and *multitrait* rating scales. These are less reflective of general theories than the holistic and analytic scales, but are more influenced by an analysis of task demands and by consideration of educational objectives and local needs and constraints. This second approach to rating language performance has responded to some of the more obvious limitations of the holistic and analytic approach. For example, congruence with instructional objectives provides opportunity for positive washback effects. The two approaches to rating scales can be illustrated in reference to assessing a learner's writing.

A *holistic* scale of writing ability would describe a number of levels of overall writing ability which are intended to be applicable to any sample of writing. Reference might be made to organization, originality, mechanics, vocabulary range and accuracy, coherence, cohesion, grammatical accuracy, etc., but each of these aspects of performance contributes to the description of general ability levels. *Analytic* scales are also designed to be applicable to various writing samples. The difference is that there is more than one scale in order to rate separately a number of different attributes that are hypothesized to contribute to overall writing ability. The separate analytic ratings may be combined to provide a

measure of general ability, but they need not be. Holistic and analytic scales are often constructed on the assumption of a hierarchy of levels in which a skill that characterizes one level is assumed to be fully acquired before a skill at the next level can be acquired. The attributes that are cited in the scales appear to derive from contemporary theories of proficiency.

Primary trait and *multitrait* scales, in contrast to the holistic and analytic rating scales, are designed to indicate how well an individual satisfies the requirements of a given task. Cohen, for example, discusses a primary trait rating scale designed to assess responses to the prompt:

> Some people believe that a woman's place is in the home. Others do not. Take ONE side of the issue. Write an essay in which you state your position and defend it.
>
> (Lloyd-Jones 1977:60 in Cohen 1994:321)

The primary trait levels all indicate (1) whether the writer has, in fact, taken a position, (2) the number of reasons given, and (3) whether or not each of the reasons is elaborated.

Multitrait scales are similar to analytic scales in that a performance receives more than one score. They are similar to primary trait scales in that each score scale is specific to the task. For example, a writing task may direct students to:

> Read the following article entitled "The appropriate steps to take when preparing for a final exam," summarize it, and criticize its content, relating your arguments to your personal experience.

This might be scored on two traits, expressing the main ideas of the article, and quality of argumentation. Two score levels for each of these five level scales are shown in Table 1.

Most rating scales, whether they be holistic, analytic, primary or multitrait scales, take the "standard" form as described by Underhill (1987)

> ...a series of short descriptions of different levels of language ability. Its purpose is to describe briefly what the typical learner at each level can do, so that it is easier for the assessor to decide what level or score to give each learner in a test. The rating scale therefore offers the assessor a series of prepared descriptions, and she then picks the one which best fits each learner.
>
> (Underhill 1987: 98)

Score Level	Trait	
	Main ideas	Argumentation
5	All main ideas of the article are included and expressed clearly.	Two or more relevant arguments are well presented and effectively related to the writer's experience.
.	.	.
.	.	.
.	.	.
1	Main ideas of the article are difficult to identify.	Arguments, if present, are inadequately presented or may be irrelevant. Relation to writer's experience may be unclear.

Table 1: Example of Multitrait Scoring Scale

At present, there is considerable concern about the validity and reliability of rating scales that conform to this description. The lack of direction for testers may account in large measure for current criticism of language rating scales[2]. Matthews (1990) claims that criteria for scoring are often arbitrary and inconsistent. She further questions both the reliability and validity of rating scales and concludes that other types of test design may be necessary. Fulcher (1987) is concerned that recent scales reflect what theorists think happens in communication situations, and not what actually happens, that is, that they are based on theory without adequate empirical support. He suggests a need for data-based criteria for evaluating performance.

Looking at different, possibly innovative, procedures for the development of rating scales can help address the issue of reliability and validity. It has been stated repeatedly in recent language testing 'state of the art' papers (Alderson 1991; Bachman 1991; Skehan 1988, 1989, 1991), that new avenues for assessing communicative abilities more successfully must continue to be explored. Criteria for evaluating performance more relevant to language development and use are needed. Tasks and scoring procedures need to be designed to accommodate the

naturalness of language use (Skehan 1991). Assessing such language effectively is more demanding than objective scoring. Upshur & Turner (1995) have summarized the limitations of standard language rating procedures.

Accounting for Reliability and Validity Problems

There are sound reasons to account for the difficulties in achieving high reliability and validity with standard rating scales. In large measure these have to do with problems of mapping a large number of qualitative differences into unidimensional metric scales. The conceptual classes that constitute the levels of a typical rating scale are often described monothetically. That is, none of the descriptors for one level are found in the descriptions of other levels. However, empirical cases that are rated are almost always polythetic; they exhibit a combination of characteristics that are described at different levels. This is an inevitable consequence of a multidimensional property space, imperfect intercorrelations among variables, and the constraints of a single scale. For example, writing ability might be conceptualized in terms of only two variables (although in practice a greater number would be more likely): *form* and *content*. If four levels of each variable are described, a property space of sixteen classes is defined. That is, it is theoretically possible for writers to have any level of formal ability in combination with any level of content ability. Yet a four level rating scale of writing ability would generally combine equivalent levels of each of the two variables. (Slightly more sophisticated procedures for reducing property spaces may be employed. These may introduce polythetic classes, but do not generally solve the problem of classifying writing samples.) Writing samples from any of the dozen other classes of the property space must be mapped into the four "ideal classes" according to a rater's own view of which is "closest" to the sample being rated. In fact, however, current theories of natural language abilities posit a large number of variables (see e.g. Bachman 1990 and included references). For example, if only half a dozen variables are considered significant to some ability, and if only five levels of each variable are described, the resulting property space describes 15,625 (i.e.,56) different classes of writing ability. The problem of reducing this number of classes to eight or ten levels of ability is not addressed by any current theory of language proficiency. Instead, raters must solve this

problem in some *ad hoc* manner when they attempt to use scales describing complex abilities. It is not surprising, therefore, that ratings with such scales are as inconsistent as they are.

One possible way to resolve the problem of rating complex abilities is to rate the significant variables independently and then to combine the results of the separate ratings mathematically. The validity of such combinations depends, of course, upon how well the mathematical model employed corresponds to the way the variables actually influence the abilities. For example, a simple additive combination assumes that all variables are of equal importance throughout the range of ability, that any variable can compensate for any other, that all variables are measured on a common interval scale, that relations among the ability and variables are linear, and that there are no interactions among variables. Less constrained models are available, but well-motivated grounds for selecting among them are lacking.

Given this lack of *theoretical* direction for rating scale construction, it is natural that traditional proficiency scale development has been criticized for its lack of empirical support and lack of established sets of principles for construction purposes (Brindley & Nunan 1992, in McKay 1995). Increasingly, we witness a need for the development and use of empirically-based rating scales, particularly to meet the needs of educational programs (Chalhoub-Deville 1995; Shohamy 1992). Chalhoub-Deville (1995:28) argues "for a research approach that derives scales empirically according to the given tests and audiences, and according to the purpose of assessment."

In the remainder of this paper we discuss an alternative method to scale construction. We have attempted, in developing an explicit procedure for creating rating scales, to ensure that the rating process demanded of raters would be similar to that employed by "experts"[3]. That is, the scale development procedure would not lead to descriptors for scale levels as is the typical format. Rather, the procedure should end with a hierarchical sequence of attribute checks. In other words, development does not begin with a theory of ability levels; rather a description of levels emerges from examination of a sample of actual task performances. In form, these scales require the rater to make a series of binary choices about features of student performance that define boundaries between score levels. The scale is **empirically** derived, requires **binary** choices

by raters, and defines the **boundaries** between score levels (EBB). A scale is developed for a single task and a single population[4]. An example of an EBB scale is shown in Figure 1. This is an imaginary scale for rating presentations at a professional meeting that has been devised for illustration only. The scale has four score levels. The meaning of each level is described on the right side of the figure. The rating process, a sequence of questions about the application file, is described on the left.

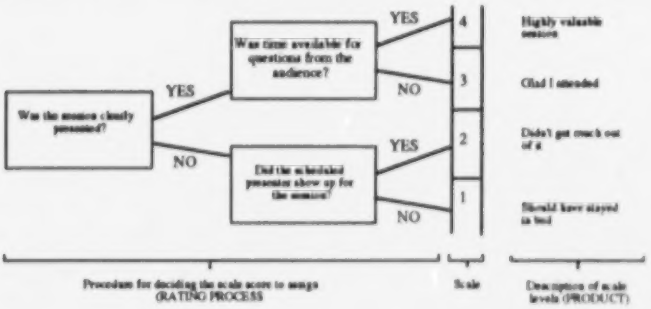


Figure 1: Form of EBB scales

The EBB scales and the scale development procedure have been designed to incorporate the following conditions and constraints:

1. Let actual performances tell what elements of the property space actually occur;
2. Do not assume what variables are important at different levels;
3. Let scale include only as many discriminable levels as raters can use reliably;
4. Assure that all levels are used;
5. Procedures for constructing scales are explicit;
6. Scoring is efficient, both in training time and rating time;
7. Incorporate knowledge and procedures followed by experts.

Procedure for EBB Scale Development

EBB scales are developed by a team of four to eight members who are familiar with the aims of assessment, the learners to be assessed, and the

BEST COPY AVAILABLE

tasks that have been used to elicit language performances from those learners. The scale results from the team's accomplishment of five tasks that constitute the development procedure. Before actual scale making can take place, however, some preliminary work needs to be accomplished. First of all, the team members need to define the trait, construct or attribute that they will be focusing on. What is it that they are to be rating? Then, taking that definition, a sample set of actual performances from a specific task and learner population must be selected. This is done by holistically identifying the range of ability exhibited in a sample, and then selecting performances that represent, in roughly equal intervals, the full range of ability[5]. The sample set should consist of *at least* eight to ten performances. It is through examination of this sample set that a scoring procedure is devised in order to assign score levels to performances. The five tasks to be performed by a team developing a rating procedure are described below.

Task 1

The first task in EBB scale development is to be done individually by each team member (see Appendix 1)[6]. Keeping in mind the trait being evaluated, each member divides the sample set into two groups: the lower half and the upper half. The purpose of this task is to familiarize the team members with the details of the performances in the sample and to have them consider, albeit implicitly, the characteristics of different levels of ability.

Task 2

The second task is composed of two parts (see Appendix 2). First, as a group the team members discuss their choices and reach consensus on which performances belong in the lower and upper halves. Second, the team decides upon a *yes-no* criterial question that allows them to classify the performances as either "lower half" or "upper half" according to the trait being evaluated. The primary purpose of this task is to identify the most salient attribute (or combination of attributes) that can serve as an indicator of the boundary between the two central levels of the rating scale.

The yes-no question formulated at this step of scale development will be applied to every performance that is rated by the developed scale. This is illustrated by the leftmost question in Figure 1.

Task 3

The third task has three parts (see Appendix 3). The objective is to determine how many score levels are appropriate for this particular population performing this particular task. First, the team members decide whether they will work with the upper-half samples or the lower-half samples. Ordinarily this would be the half within which most crucial decisions would be made. For example, a course passing level would usually be in the lower half. That is, more than half of the students in a course would be expected to pass. In contrast, the cutoff level for a highly selective selection test would probably be in the upper half.

Second, each individual team member ranks the performances from weakest to strongest. Some may be judged to be equal. Third, the team members compare their decisions and look to see how many clusters they agree upon. This will lead to the number of score levels appropriate for the half scale. The other half of the scale will use the same number of levels.

Task 4

The fourth task is to formulate the remaining yes-no criterial questions that sort the performances into the score levels (see Appendix 4)[7]. Successive subsamples are divided, and critical questions for making the divisions are formulated. The procedure is similar to that employed in Task 2.

Task 5

Tasks 1-4 develop a procedure to guide raters as they score performances. Interpretation of each score level is not evident, however. The objective of Task 5 is to write descriptions of each score level in order to provide a more complete picture of the trait being evaluated (see Appendix 5). These level descriptions are especially useful to such score recipients as parents, tutors, etc.

A CASE STUDY: AN APPLICATION OF EBB SCALING

Setting and problem

The procedure for developing EBB scales was implemented in an educational setting through a collaborative project involving university

based researchers and ESL teachers working in a large urban school board in the province of Quebec. The board is part of the French school system where English as a second language (ESL) is a required subject starting in Grade 4. At the end of Grade 6, all students are placed in secondary schools. Students are streamed into ESL classes according to their ESL marks at the end of Grade 6.

The ESL curricular advisor from a local school board brought to the researchers' attention that the board wanted to develop a standardized test of speaking ability for fast, efficient teacher administration to Grade 6 ESL students. The ESL Grade 6 teachers were using their own speaking ability testing and trying to adapt the results to a standard-type scale put forth by the Ministry of Education. At the time, there was concern that speaking performance assessments given in one school by one teacher might not be comparable to those given by another teacher in another school. Some problems identified by the teachers were: inconsistency in type of task being used, lack of agreement among teachers on the meaning of scale descriptors, only a portion of the scale was applicable to their classroom situation, important distinctions among students were not reflected in the scale, and the scale descriptors did not conform to the teachers' learning objectives. For these reasons the board wanted to equate standards across all schools so the results could be used as one indicator for secondary school ESL class placement.

The project was therefore set up with the intent that the researchers could assist the school board in test development, EBB scaling and training of the teachers as raters (judges). From a research point of view, the development of task- and population-specific rating scales and effects of their use could be examined.

TEST DEVELOPMENT

From the outset, certain points were agreed upon among the participants: (1) that this was to be a collaborative endeavor with university researchers, the school board and teachers, and (2) that the new test would generate actual speech (i.e., natural speech) from the students for evaluation. The participants in the development, administration and scoring of the tests were the curriculum advisor, 12 elementary school teachers and two researchers.

There were two groups of participants involved in this project: (1) 12 elementary school teachers who served both as test developers and raters, and (2) 255 Grade 6 ESL learners. The students were in a school board which included a variety of socio-economic neighborhoods. Since Grade 4, the students had followed the regular ESL program of 120 minutes per week. Their Grade 6 program was based on communicative language teaching. Teacher participants were volunteers. Student participants were enrolled in the classes of the teacher participants. Standard practices for use of human subjects were followed.

Instruments

This project employed two measures of speaking ability developed especially for the investigation of this population. A measure is defined here as a speech elicitation task and a rating scale. The development process involved the 12 teachers and the two researchers. The EBB scaling procedure was followed. Development of measures included a sequence of activities: identifying the trait to be evaluated, task selection, developing administrative procedures (tryout and revise), and developing the rating scales. On-going communication and working sessions among the participants enhanced the endeavor.

Elicitation tasks

The Ministry of Education prescribes that a student finishing Grade 6 "...should be able to use English as a means of communication in situations which pertain to his/her needs and to the interests characteristic of his/her age" (Ministère de l'Éducation 1984). For the purposes of this assessment, that objective was operationalized as: "When performing this task, how well does the student make him/herself understood?" The development group agreed that this ability is highly complex, and that many skills contribute to the ability to make oneself understood. To develop this ability a student must attain many different learning objectives. In rating the ability, it was decided to first attend to the overall effect of a student's speech. Indicators of different levels of ability would be discovered during EBB scale development.

A half-day meeting was held with the development group to identify or develop elicitation tasks for the measures. The participating

teachers brought suggestions for speaking tasks that would be: productive, feasible and familiar to the students. Initial suggestions were reviewed and revised. Two tasks were agreed upon and developed for use in the project. The two tasks were: Story Retell and Audio-Pal.

Story Retell - The students watch a 2.5 minute video. They are instructed to draw a picture after viewing the video to help them remember the story. They are asked to save their pictures. Individually they go to a quiet place where they find a tape recorder and an instruction sheet. They retell the story in their own words using their pictures.

> The video is entitled *Arnold of the Ducks* (Columbia Broadcasting System 1990). The cartoon video shows Arnold, a child about one and a half years old, playing in a wading pool in his backyard. Arnold's parents, busy with household chores, assign the family dog to keep an eye on the boy. The dog goes to sleep, so Arnold is left unattended. Soon a near-sighted pelican flies over and, mistaking the child for a fish, swoops down, picks him up and flies off. The parents, alerted by the barking dog, discover that their son is gone but fail to see him happily gazing out from the pelican's beak as the bird soars off over Arnold's home and town.

Audio-Pal - Students go individually to a recording site. They are informed that English-speaking exchange students of the same grade from other provinces are coming to live and attend school with them for a month. The student is instructed to compose a letter in the form of a tape recording to that exchange student. They are guided to talk about things like: "yourself and family, your hobbies and interests, things you might do together, what school is like".

The elicitation tasks were piloted and revised. With the help of teacher feedback, administrative procedures were developed for this trial and then fine tuned for the final administration: general logistics for testing, instructions for teachers, instructions for students, materials and equipment, and follow-up procedures.

Rating Scales

Rating scale development took place after the teachers had administered the elicitation tasks to a sample of 36 Grade 6 learners from across the school board and recorded their performances. The two researchers performed the preliminary work of identifying sample sets from the data.

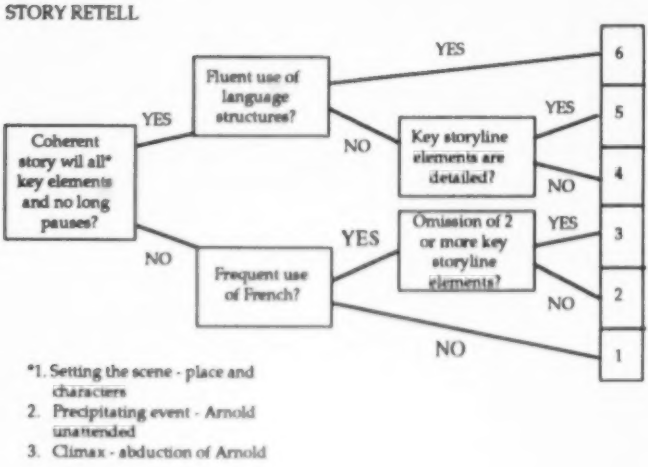


Figure 2: Rating scale procedure for Story Retell

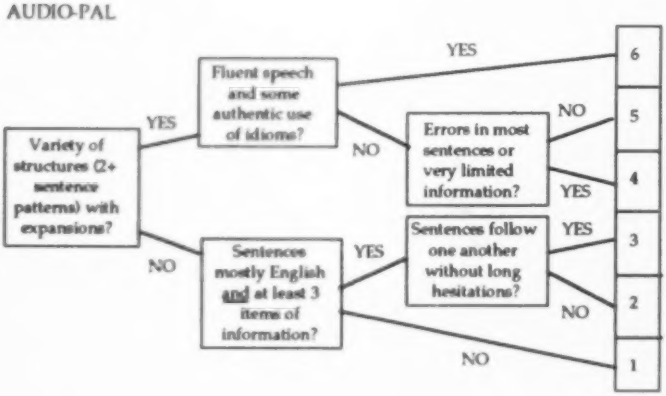


Figure 3: Rating scale procedure for Audio Pal

A sample set of 12 performances was identified for each task (i.e., one sample set from Story Retell responses and one sample set from Audio-Pal responses).

Two rating scales were empirically developed from the sample sets: one scale specific to the Story Retell task, and the other specific to the Audio-Pal task. Six teachers and one of the researchers developed one scale, while the other six teachers and second researcher developed the other scale. This took place over a one-day working session. The criteria for scale development were: task specific, empirical, teacher-made, user-friendly, and fast to use. When listening to the student responses in the sample sets, the teachers focused on the overall effect of a student's speech. They kept in mind the agreed upon assessment question: "When performing this task, how well do the students make themselves understood?"

During this one-day working session, the teachers were able to complete Tasks 1 through 4 (see Appendices 1-4). This demanded continual intensive focus in order to succeed. Even though the teachers all taught a common program within the same school board, the scale development process generated much discussion, which demonstrated individual teaching philosophies and ideas about effective oral language. Over the course of the day, however, consensus was reached at each step of the scale development procedure. The final scales are shown in Figures 2 and 3.

Due to time constraints, Task 5 (the written descriptions of each score level for interpretation purposes), was not completed during the working session. Teachers were requested to complete the task and send it to the researchers. Six teachers responded, but only one of the responses was usable. More direct guidance was needed on Task 5. The procedure called for listening to performance samples in each level and then deriving a reasonably complete description of the common characteristics and the variations. An example is shown in Figure 4. This was not produced by the teachers, but by the authors for illustration only.

Scoring Procedure

Near the end of the school year, teachers used the two tasks to test their students. A scoring day session was set up soon thereafter. All of the teachers brought in the tape recordings of their students. The day began

Level 6

Communication is generally effective; the task is successfully performed. Although they are all effective, learners show considerable diversity in the means utilized to attain this standard. Fluency ranges from somewhat "choppy" to a completely natural flow that appears almost native-like. Normal cohesion markers may be missing, but the organization of the discourse is sufficient to provide adequate coherence. Grammatical accuracy is almost complete for some speakers except for occasional local fossilizations. In no cases are there gross syntactic errors that impede comprehension because they co-occur with limitations in other aspects of speech. In all cases, speakers utilize a variety of sentence patterns. Vocabulary is used accurately, for some speakers idiomatically. Many speakers sound somewhat stilted, however. The most able speakers at this level seem fully at ease in their use of English. The least able speakers evidence limitations in some aspects of their speech, but these limitations are well compensated for by strength in other aspects or by skillful use of communication strategies.

Level 5

Communication is fairly effective; task is accomplished, but with some limitations. All speakers address the topic and provide relevant information. The flow of speech is never native-like. For some speakers it is only a slight unnaturalness; for the least fluent at this level, unnatural pauses and hesitations in the speech are clearly evident and sometimes intrusive. These seem to be usually related to search for vocabulary. Discourse is coherent for most speakers, achieved by most without a natural-seeming use of formal markers. For the least able speakers at this level, lack of organization and markers may make the function of an utterance within the discourse obscure or ambiguous. Few grammar errors are made and a variety of forms are employed with expansions. Vocabulary shows breadth and is, for most speakers accurate. For some speakers general terms are used in place of more appropriate, specific vocabulary. Many of the speakers at this level exhibit a range of communication strategies without recourse to use of the first language.

Figure 4: Example of level descriptors, Levels 5 and 6, Audio-Pal

by reviewing the two rating scales that had been developed by the teachers. Half the group had developed one scale and the other half the second scale. Therefore, it was necessary to familiarize the teachers with the scale they had not developed as well as to review the scale they had developed. Time was also taken to practice with each scale. For the rest of the day, teachers rated their own students and the students of two other teachers. In other words, each student recording was rated independently by three of the participating teachers. Each teacher had approximately 67 performances to rate. They managed to rate approximately 15 to 20 performances an hour.

Teacher Feedback

Throughout the day of scoring, and at a later feedback session, the teachers expressed the benefit they received from participating in this project. The general opinion was that the process had helped demystify testing and evaluation for them. After being test developers themselves, they felt more comfortable with the evaluation process. They took pride in, and ownership of their own product. They felt their product had been successful due to a team effort that had required much discussion, exchange, and finally a consensus. In addition, it is interesting to note that this experience caused a positive washback effect on their teaching methodology. Teachers began using more tasks that involved speaking into a tape recorder. The students enjoyed the independent practice and the teachers expressed pleasure at having extended discourse samples of their students. As one teacher expressed, "It really gave me an idea of their abilities in English."

The school board, ESL curriculum advisor and teachers made the decision to continue with this procedure for summative assessment of Grade 6 speaking ability. The researchers were invited back to guide them in the initial stages of their second test development project.

CONCLUSIONS

The primary conclusions we can draw from this attempt at construction and application of EBB scaling in the field are related to our success in satisfying the conditions and constraints we established for ourselves (see the list on page 61, above).

1. The empirical procedure for deriving sorting rules assures that these are related to actual features found in the sample utilized for scale construction. These were robust enough that no raters reported difficulty scoring any speech samples with the scales.
2. The development procedure allowed important variables to "emerge". The importance of this constraint was shown in two ways. First, crucial variables differ as one changes levels. For example, the distinction between levels 4 and 5 on the Story Retell scale is based upon provision of details; The distinction between levels 2 and 3 is based upon the inclusion of basic story elements (see Figure 2). Second, different variables are crucial at same level boundaries when rating different tasks. For example, in comparison to the Story Retell dependence upon inclusion of major content to distinguish level 2 and 3, Audio-Pal depends upon fluency.
3. The third task of the development procedure assures that the scale does not require scaling into more levels than the teachers, as a group, can perceive. That is, when there is not general agreement that one speech sample exceeds another in quality, no attempt is made to make score distinctions between them.
4. All score levels were used. Assuming a somewhat random distribution of speaking ability among students, relatively few students should be rated 1 or 6; the most frequently assigned rating should be 3 or 4. In fact, for the Story Retell 4% and 7% of the ratings were 1 and 6 respectively. Corresponding percentages for Audio Pal were 3% and 9%. All other score levels were assigned for at least 10% of the total number of ratings for a test. The most common rating was level 4, (35%) for Audio Pal, (27%) for Story Retell.
5. The procedures devised for constructing EBB scales are embodied in tasks 1-5 (see appendices below). These were sufficiently explicit to allow a group of teachers who had never made such scales before to do so with minimal guidance. The teachers agreed that they would be able to follow the procedure without guidance in a future attempt.
6. Scoring of speaking ability was quite efficient. All of the raters had participated in the construction of one scale for one of the two test tasks. This work, approximately five hours, familiarized them

with EBB type scales. With that background, training time for using the scale they had not worked on was one hour. They were then able to rate 15-20 tests per hour. This figure underestimates efficiency, however. A lot of time was spent waiting for tapes, etc., requirements of the research aspects of the rating session. Few raters actually had to listen to tapes more than one time. Thus potential for efficiency in operational testing is very high. Time would probably be cut in half if raters could rate all tapes for one class at a time, then have another rater rate the same set of tapes at leisure. The reliability of ratings was quite satisfactory. A Facets analysis of results yielded a reliability coefficient of .85. This compares very favorably with reliability attained with much more time consuming testing and rating of speaking ability (Henning & Cascallar 1992 ; Shohamy 1983; Upshur & Turner 1995). The reliability estimate for EBB scaling is based upon averaging the ratings of two or three raters. Reliability would be lower if each student was rated by a single rater. We found that raters differed in their severity, but not greatly so in comparison with the relatively wide range of student ability. We believe, however, that training time for raters could be profitably increased from the one hour that we provided.

Because EBB scales are task-specific, they are not efficient for daily, formative assessment. The time needed for sampling student performances and organizing scale development make EBB scaling more appropriate for occasional high-stakes testing.

7. Rating efficiency suggests that raters employed the "expert" procedures that EBB scales were designed to favor. No systematic investigation of this question was undertaken, but insofar as volunteered reports from raters indicate, the procedures were being used.

Our project did not include a formal study of the validity of the two tests that were developed. Test validity is not purely a function of scale type or rating procedure. Nonetheless these will affect the rating of a test. Therefore, although a focused investigation of validity was not made, some speculations are possible. These support claims for content validity of the tests. First of all, test tasks were selected by teachers to reflect activities that their students were expected to perform in the course of

their language learning. In a communicative course these activities not only represent instructional techniques, but also define aims of the course.

A second validity claim relates more directly to scale development and the rating procedure. It was teachers of the course who made the scoring rules based upon actual speech samples. These were, therefore, necessarily relevant to student performance. Furthermore, the features of speech that were incorporated into the rules would be features that were salient to the teachers, most likely because they were concerns in instruction. That is, the scales incorporate requirements for scoring on the basis of what teachers value and are able to recognize. Because scaling procedures and level descriptors are not derived from an explicit theory of second language performance, the tests are more relevant to educational practice than to theory. This is especially so when program aims and practices may diverge from theoretical prescriptions. For this reason these scales are not recommended for testing student attainment of abilities that are not included in program aims and objectives.

In summary, we found EBB scaling to be feasible and efficient. These scales are not recommended for all instances in which performance ratings are wanted. However, for relatively high-stakes testing when particular theoretical constructs need not be investigated, the scales provide some clear advantages over the traditional approaches to performance rating.

NOTES

1. We wish to thank Normand Lapointe, General Director, the curriculum advisor, Anna Belle Barakett, and the following teachers from the Commission scolaire du Goéland for their cooperation in the scale development project: Henny Blom, Yvan Bédard, Lucie Couture, Suzanne Daoust, Susan Edwards-Girard, Andrée Lavarière, Kim O'Halloran, Maria Papandrea, Mary Steward-Wood, Louise Triggs, Anne Turpin-Lajoie and Marina Walter. Thank you also to Betty Chow, our research assistant. This project was partially funded by the Social Sciences Research Grant Committee, McGill University.
2. See Alderson et al. (1995:111-113) however, for directions about ways to use existing scales.

3. In the hoary past the second author asked a half dozen experienced composition raters to observe and reflect upon the procedures they followed while reading and scoring. All used some sort of sequential, hierarchical procedure. That is, they first checked for one quality of the compositions, generally something apparent upon a fast skimming of the text. An initial tentative score or a range of scores was decided on the basis of this first check. Then the "rough-graded" composition was checked for another quality in order to narrow further the possible score range. This process was continued until a final scale score was decided.
4. In fact, as Michel Laurier has pointed out (pc), we are proposing a decision algorithm, a means for assigning scale values to language performances. As shown in Figure 1, the scale is an ordinal sequence of numbers.
5. In the past, this task has been performed by someone who is familiar with the test development project but does not serve on the scale development team.
6. The actual task descriptions in this paper were devised for use with a sample of twelve speech samples.
7. The Task 4 description is one that is used when a six level rating scale has been determined. Descriptions that show a different question tree are available for four level and eight level scales.

REFERENCES

Alderson, J.C. (1991) Language testing in the 1990s: How far have we come? How much further have we to go? In S. Anivan (ed.) *Current Developments in Language Testing*. Singapore, SEAMEO Regional Language Centre.

Alderson, J. C., C. Clapham and D. Wall (1995) *Language test construction and evaluation*. Cambridge, Cambridge University Press.

Bachman, L.F. (1990) *Fundamental considerations in language testing*. Oxford, Oxford University Press.

Bachman, L.F. (1991) What does language testing have to offer? *TESOL Quarterly* 25:671-704.

Chalhoub-Deville, M. (1995) Deriving oral assessment scales across different tests and rater groups. *Language Testing* 12:16-33.

Cohen, A. (1994) *Assessing language ability in the classroom.* 2nd Edition. Boston, Heinle and Heinle.

Columbia Broadcasting System (Producer) (1990) *Arnold of the ducks* [Videotape].

Fulcher, G. (1987) Tests of oral performance: The need for data-based criteria. *EST Journal* 41:287-291.

Henning, G. & Cascallar, E. (1992) *A preliminary study of the nature of communicative competence.* TOEFL Research Reports, Report 36. Princeton, NJ, Educational Testing Service.

Matthews, M (1990) The measurement of productive skills: Doubts concerning the assessment criteria of certain public examinations. *ELT Journal* 44:117-121.

McKay, P. (1995) Developing ESL proficiency descriptions for the school context: The NLLIA bandscales. In G. Brindley (ed.) *Language assessment in action.* (Research Series 8). Sydney, NCELTR, Macquarie University.

Ministère de l'Éducation. (1984) *Elementary school curriculum: English as a Second Lanugage* (#16-2204 A). Québec, QC, Gouvernement du Québec, Ministère de l'Éducation.

Shohamy, E. (1983) Rater reliability of the oral interview speaking test. *Foreign Language Annals* 16:219-222.

Shohamy, E. (1992) New modes of assessment: The connection between testing and learning. In E. Shohamy and A.R. Walton (eds) *Language assessment for feedback: Testing and other strategies.* Dubuque, IA, Kendall/Hunt.

Shohamy, E. (1995) Performance assessment in language testing. *Annual Review of Applied Linguistics* 15,:88-211.

Skehan, P. (1988) State of the art article: Language testing I. *Language Teaching* 21:211-221.

Skehan, P. (1989) State of the art article: Language testing II. *Language Teaching* 22:1-13.

Skehan, P. (1991) Progress in language testing: The 1990s. in J.C. Alderson and B. North (eds) *Language testing in the 1990s: The communicative legacy.* London, Modern English Publications and The British Council.

Underhill, N. (1987) Testing spoken language: A handbook of oral testing techniques. Cambridge, Cambridge University Press.

Upshur, J. A. & Turner, C. E. (1995) Constructing rating scales for second language tests. *ELT Journal* 49,1:3-12.

APPENDICES

Tasks for the development of rating procedures

Appendix 1 Task 1

TASK 1

The first part of this task is to be done independently, without discussion or consultation. It is important that each teacher's judgement is independent.

Divide the twelve speech samples into two groups: the better six and the poorer six. Do not try to rank all of the samples; just decide which ones make up the upper half and which ones the lower half. Record the numbers of the samples below.

UPPER SIX: ___, ___, ___, ___, ___, ___
LOWER SIX: ___, ___, ___, ___, ___, ___

In the space below you can make brief notes on each of the speech samples.

1	
2	
. . .	. . .
12	

Appendix 2 Task 2

TASK 2

The object of this task is to create the first step in the process for assigning scores to speech samples elicited by this test of speaking ability. There are two parts to the task:

First you will have to come to agreement within your group about which samples are to be considered upper and which are to be considered lower.

Next you will have to decide on a yes-no question that a rater would ask about a speech sample. One answer would sort the sample into the upper score range, the other would sort the sample into the lower range. (Either a "Yes" or a "No" could be used for upper scores depending on whether the question asked about positive or negative attributes of the samples.)

PART 1 (The following suggestions may be helpful; you may prefer other procedures.)

- One member of the group should act as recorder.
- First find out whether there are any samples about which there is consensus.
- Discuss the samples for which there is not consensus. Teachers should each explain the basis for rankings that differ from those of other teachers. Sometimes this will lead to resolution of differences. If consensus cannot be achieved, a vote can resolve the impasse. All individual opinions are equally valid, but compromise may be necessary in order to develop a measure that can be uniformly applied.

Record the agreed division of samples below:

UPPER: ___, ___, ___, ___, ___, ___

LOWER: ___, ___, ___, ___, ___, ___

PART 2

- Keep your question as brief as possible (although you may have to amplify in a note).
- Your question should relate to an actual, observed difference between the upper and lower samples.
- Useful questions for this step are often related to the ability one is assessing, in this case the ability to make oneself understood.
- Reference to the earlier discussions about samples that caused disagreement may help you to find a usable question.
- The best questions refer to things that are relatively easy for teachers to recognize.

Record your team's question below:

Q-1	

BEST COPY AVAILABLE

Appendix 3 Task 3

TASK 3

The object of this task is to learn how many score levels are appropriate for these students taking this test. For some kinds of performance only two score levels are possible. Many multiple-choice or short-answer test items are counted either right or wrong, with one score given for a right answer and another, usually 0, for a wrong answer. Other test performances may be rated on a scale with many levels. Some teachers will, for example, rate student compositions on a 100 point scale even though it has been demonstrated many times that teachers cannot hold in memory standards for so many degrees of difference. In general, ten levels is about as many as one could ever hope for. Depending upon the kind of performance being rated and the differences among the people performing, two to eight levels is more reasonable. For speech samples of approximately a minute by students in one grade, it should be relatively easy to distinguish four levels, six is possible, and even eight is conceivable although not too likely.

The procedure for deciding on the number of levels possible is to compare the rankings given by a number of judges to a number of speech samples. When many judges disagree about the relative ranking of two speech samples, this indicates a single level. When the judges are virtually unanimous, this suggests that the speech samples belong at different levels.

PART 1

Decide as a group whether you will rank the six upper speech samples or the six lower.

PART 2

Independently rank the six samples from weakest to best. If you think that any set of samples are indistinguishable in quality, order them arbitrarily and insert an equal sign (=) to show that they are the same. Use the sample numbers and record your rankings below:

WEAKEST ___ ___ ___ ___ ___ ___ BEST

PART 3

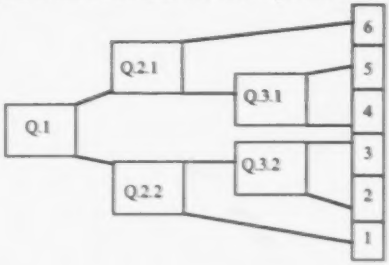
Compare the individual rankings to see how many clusters of ability the group agrees upon. These clusters of ability will be score levels on the scale you develop. Carolyn or Jack can help out with this part.

(Note: In the unlikely event that no clusters are agreed upon, we will have to divide the set of samples into at least two score levels. Or if there is complete unanimity on six different performance levels, the number of score levels will have to be reduced to three or four.)
The same number of score levels will be defined for the other half of the scale.

Appendix 4 Task 4

TASK 4

The object of this task is to complete the set of questions needed to sort speech samples into the score levels of your scale. The number of levels you found to be appropriate in the previous task determines the number of questions (the decision model) you will need. The procedure for creating questions is like the procedure used for creating the upper- lower-half question in Task 2.



Appendix 5 Task 5

TASK 5

The object of this task is to provide descriptions of the speaking abilities of students at each of the score levels of your scale. Through the previous tasks we have created a set of questions that guide raters in their assignment of scores. Because of their particular purpose, the set of questions does not provide a full picture of the different levels of speaking ability. The description of score levels will provide an interpretation for scores on this test.

Before beginning your descriptions, select one member of the group to act as recorder.

The task is best performed with consultation among all group members.
First sort the speech samples into score levels. For each score level review the notes made about each speech sample and listen again to the samples to derive a reasonably complete description of the common characteristics (positive and negative) and the variation in the performances at that level.
In the space below, write descriptions for the score levels used in your scale.

Score	Description
6	
.	.
.	.
.	.
1	

ADMINISTERING A PERFORMANCE TEST IN FIJI HINDI

Nikhat Shameem
University of Auckland
John Read
Victoria University of Wellington

ABSTRACT

As part of a research study on language maintenance and shift in the Indo-Fijian community in Wellington, New Zealand, a performance test was developed to assess the speaking and listening skills in Fiji Hindi of a sample of the Indo-Fijian teenagers. The design of the test needed to take account of the fact that Fiji Hindi is a preliterate vernacular language with no role in education and an ambivalent status within its own speech community. The test consisted of three main parts: a naturalistic conversation, two structured speaking tasks and a structured listening task. This paper focuses on some facets of the test administration, including the decision to administer it in the test-takers' homes; the influence of various personal attributes of the interviewer; the ways of dealing with the lack of a script for Fiji Hindi; and the issue of live versus tape-based assessment of the test-takers' performance. Both the test-takers and an independent rater provided feedback on the test that was generally very positive. The paper concludes with a discussion of various factors that may have influenced the reliability and validity of this somewhat unconventional language test.

INTRODUCTION

In language testing, as in any other form of educational measurement, the procedures involved in administering the test need to be carefully worked out by the test developers and closely followed by invigilators and examiners. Any omissions or inconsistencies will influence the reliability, and thereby the validity, of the test results. Bachman's (1990:111-157) framework of test method facets represents a comprehensive checklist of the variables which can affect test administration. Considered in terms of Bachman's framework, a speaking test poses special challenges for test design, particularly when it is individually administered over a period of time by examiners who are trained to tailor the test to the interests and proficiency levels of individual

test-takers. There are various kinds of compromise that have to be made in an effort to achieve an optimum balance among the somewhat competing requirements of reliability, validity and practicality.

If test-takers are interviewed one by one, how is it possible to achieve the kind of consistency in testing conditions that is necessary for reliable assessement? If the test is carefully structured in the interests of reliability, will the examiner be able to elicit "natural" speech from the test-taker? If two or more raters are assessing test-taker performance, how can we ensure that they are applying the same rating criteria?

The speaking test to be discussed here presented challenges of this kind — and a number of others besides. The type of test that language testers are most commonly engaged in developing involves a standard language with a well-established role in education as well as in the society at large. In that situation, the test-takers can be assumed to have an incentive to perform to the best of their ability; in fact, major tests have a significant impact on their prospects for further education or employment. By contrast, tests that are used purely for research purposes do not have the same kind of motivating effect on the test-takers. The problem was compounded in the present case by the fact that the language to be assessed is a non-standard one with no accepted written form, no role in education and uncertain status within its own speech community.

THE CONTEXT

The language that was the focus of the test to be discussed here is Fiji Hindi, the mother tongue of most of the Indian population of Fiji. It has evolved from a pidginized lingua franca that arose among speakers of many different Indian languages and dialects on the plantations in Fiji during the period of indentured labour from 1879 to 1920. For various reasons, about half of the Indian labourers remained in Fiji once their period of indenture was over. The language creolized as it was acquired by the first generation of Indian children to be born in Fiji and developed its own character, which makes it distinct from any variety of Hindi in India. Meanwhile, the Indo-Fijian population grew and prospered in their adopted country, to the point where they outnumbered the indigenous Fijians by the mid-1980s. The 1987 general election was won by a coalition that had strong support from the Indo-Fijian community. Shortly

afterwards, the Fiji military forces launched two coups with the avowed purpose of re-asserting indigenous political rights, which were perceived by many Fijians to be under threat from the new government.

In the aftermath of the coups, a substantial number of Indo-Fijians emigrated to Canada, Australia and New Zealand. By 1994 it was estimated from official figures that there were over 16,000 of them resident in New Zealand as a whole (Shameem 1995:10). In the course of the present research, a total of 62 Indo-Fijian households were identified in the Wellington region, where they form a small, dispersed minority within a highly monolingual English-speaking society. Previous studies of language maintenance and shift among language minority groups in Wellington (see, e.g., Holmes, Roberts, Verivaki and 'Aipolo 1993) have shown that, while there is a general shift to English over time, the process has occurred at a variable pace, depending on the effect of various external and internal factors on the group concerned. Research on the Indo-Fijian community presented the opportunity to study a language minority group that was newly arrived in the host society, so that a baseline could be established against which future sociolinguistic trends within the community could be judged.

The first phase of the research (Shameem 1994, 1995) involved a survey of language proficiency and use among members of the Indo-Fijian community in Wellington. The main subjects were a stratified sample of the Indo-Fijian teenagers, together with their mothers. Teenagers were chosen as the primary subjects because, although on the one hand they could mostly be assumed to be proficient in Fiji Hindi, they would also be more susceptible than their parents to pressure from their peers to speak English, especially in the school environment. The small number of Indo-Fijian teenagers at any one school was likely to mean that there was little or no possibility of using Fiji Hindi at school. In addition, as the children of immigrant families they would be subject to pressure to achieve academically in order to take advantage of the educational and economic opportunities offered by their new country; This would inevitably mean attaining a high level of competence in English. In a real sense, then, the way in which the teenagers reacted to these various factors favouring shift to English would determine the future prospects of Fiji Hindi in New Zealand.

For the first phase, the data was collected by means of a structured interview conducted with the teenagers and their mothers in the family

home. A check on the validity of the teenagers' responses could be made by comparing what they said with what their mothers reported about the teenagers' level of proficiency in, and extent of use of, Fiji Hindi and the other languages in their speech repertoire. Nevertheless, the first phase relied crucially on self-report. This indeed — along with observation — is the conventional method of data-gathering in research on language maintenance and shift in particular language communities. However, sociolinguists (e.g. Fasold 1984:113-124, 215-216; Romaine 1989:12-22) are well aware of the limitations of census returns and other survey data based on people's reports of their language proficiency or language use. Relatively few studies (see, e.g., Hakuta and D'Andrea 1992) involve any direct assessment of the subjects' language proficiency, especially if their own language is a vernacular that is not used in school. It is desirable, though, in this kind of research to be able to administer some form of oral-aural test to provide evidence of proficiency that is more direct and more comparable across subjects than that obtained through self-report.

The opportunity arose in this case to conduct a second phase of the research, in which a reduced sample of the teenagers could be given a performance test of their speaking and listening skills in Fiji Hindi. However, before the data-gathering could be carried out, a suitable test had to be developed and this meant dealing with a number of difficult issues of test design.

THE DESIGN OF THE TEST

Some of the constraints on the design of the test have already been touched on above.

- The language itself is a "preliterate" one (cf. Kloss 1968) with no established written form and no tradition of literacy associated with it.
- Fiji Hindi has no role in the education system in Fiji. The country has a policy of vernacular education in the first two years of primary school but for Indo-Fijian children that means being taught in Shudh Hindi, the standard variety of Hindi. To some extent there has been a diglossic relationship between the

two languages: Fiji Hindi is the L variety, which is used at home and for informal communication, while Shudh Hindi is the H variety, with a limited range of uses in education, broadcasting and other formal domains. However, the language situation is not stable, with English tending to take over the H role in the repertoire of Indo-Fijians.

- Since it has not had any educational role, there has been no reason until now for anyone to consider how proficiency in Fiji Hindi should be assessed. It was by no means obvious how to characterize "good" Fiji Hindi or to define what native proficiency in the language might be. This created difficulties in writing rating scales and finding anyone other than the researcher who could be regarded as competent at rating the test-takers' performances.
- Another consequence of the nature of Fiji Hindi is that Indo-Fijians have ambivalent attitudes towards it. In the first phase of the research, although they accepted it as a language in its own right, a majority of the subjects agreed that it was a "broken" language, especially when compared with Shudh Hindi. This is an attitude that is commonly expressed by Indo-Fijians in Fiji.
- Given these various other factors, it could not be assumed that the subjects would be willing to take the test or that they would be motivated to perform to the best of their ability in it.

With these constraints in mind, the test was developed broadly along the lines of the ILR Oral Proficiency Interview (OPI) and the various other speaking tests that it has inspired. There were three main components:

- a naturalistic conversation
- two structured speaking tasks
- a structured listening task

It was particularly important to enlist the teenagers' participation and put them at ease, so that they would be encouraged to perform to the best of their ability. In fact, at least in the early stages, the less it seemed like a real test, the better. Therefore, the first part of the test was intended

to simulate as closely as possible a normal conversation between two Indo-Fijians who were just getting to know each other. This meant that the test-takers were welcome to take the initiative and change the topic if they so desired. At the same time the interviewer imposed a covert structure by working unobtrusively through a series of topics that represented progressively higher levels of proficiency (from 1 to 6, see Appendix) in Fiji Hindi. The sequence began with greetings and introductions, then moved on to familiar topics such as the family, the school, Wellington, Fiji and the experience of immigration to New Zealand. At the most advanced level, the teenagers were expected to be able to discuss the issue of maintaining Fiji Hindi in New Zealand. However, the interviewer proceeded only to the point where the test-takers' "linguistic ceiling" seemed to have been reached; once there were clear indications of difficulty in expressing themselves, the interviewer moved to a topic at a more comfortable performance level for the teenager concerned.

Provided that the test-taker had demonstrated at least level 2 proficiency in Fiji Hindi, the two structured speaking tasks were then introduced. The first was based on a series of drawings depicting the steps in making an omelette and the test-taker was asked to describe the process, using the pictures as cues. The other one was a read and retell task. The test-taker was given a short story written in simple English called "Arrival" (see the Appendix for the text), which expressed the feelings of a young Indo-Fijian immigrant arriving in New Zealand for the first time. The task was to read the story and retell it in Fiji Hindi, without referring to the written text. The rationale for using these tasks was twofold. First, they contributed to content validity by broadening the range of culturally relevant topics and speaking skills covered in the test and secondly, they provided a better basis for discriminating at the higher proficiency levels.

These first two parts of the test were expected to last for about twenty minutes but with the most proficient teenagers they would often continue for a somewhat longer time.

The third main component of the test was the listening task. Although some assessment of listening ability was made on the basis of performance in the initial conversation, it was considered necessary to supplement it with a structured task to provide further evidence of the test-taker's level of comprehension. The stimulus material comprised a

recording of two excerpts from *Adhura Sapna*, a play written in Fiji Hindi by Raymond Pillai in 1990. The play has a rural Fijian setting and is steeped in Indo-Fijian culture, with the result that it requires an understanding of Fiji Hindi as spoken in the rural areas and by the older generation. The dialogue is rich in idioms, humour, nuances and cultural references. To measure comprehension, there were 20 questions yielding 36 possible answers, which could be classified into three types: trivial, local and global (following Shohamy and Inbar 1991). The test-takers could respond in Fiji Hindi, English or a mixed form. The listening task was also timed to take about twenty minutes.

Both speaking and listening were rated on separate six-point scales. The six levels had been originally defined in the form of simple rating scales for use in the earlier self-report interviews. For the performance test, more elaborate descriptions were developed to guide the interviewer in conducting the test and to specify the evidence to be considered in rating the test-takers' performance. Here, for example, is the description for Level 3 on the speaking scale:

> 3 Minimum social proficiency. Participates in basic social intercourse on familiar places and things. Still expecting simple sentence answers but evidence of wider vocabulary - related to familiar social/cultural issues. Able to use time phrases *kal, aaj, bhiyan.*. Responds with appropriate vocabulary. Knows base colours *karia, uJar, lal*. Pronunciation of borrowed/transferred English words has NZ accent. Seeks clarification appropriately. Has difficulty in responding to probe questions - hesitates with wide pauses; or repeats interlocutor's statements. Shows some ability to re-tell short story "Arrival" and describe parts of the process of making an omelette. Responds appropriately to questions requiring trivial answers in FH play extracts from *Adhura Sapna* (Questions 2, 3, 14). Basic local responses in FH to questions in play extracts.

The description draws on various kinds of evidence from the three main parts of the test. The scales were developed largely on an a priori basis in conjunction with the design of the test tasks. However, they were modified to a limited extent as a result of the small amount of trialling that was possible before the formal testing began.

The interviewer rated each test-taker at the end of the interview. In addition, all the interviews were recorded on audiotape for later re-rating by the interviewer and by a second, independent rater.

FACETS OF TEST ADMINISTRATION

The setting

A conventional language test is administered at a specific location — normally a room at an educational institution — and test-takers are expected to travel there to participate in it. In the initial stages of planning the Fiji Hindi test, we thought of doing something similar: bringing the subjects together in one setting on a particular day. The idea was to attract them to participate by organizing activities that would appeal to teenagers and at the same time would offer multiple opportunities to observe how much and how well they were able to use Fiji Hindi. The range of activities might have included role plays, structured conversations, watching excerpts from a Fiji Hindi play, and group discussion centred around reactions to the play. This idea was rejected for several reasons: it would have been complex to organize; there was no guarantee that all or most of the subjects would have attended and willingly participated; and it would have been difficult to achieve the right balance between creating an enjoyable social experience for the teenagers and having the kind of structured programme that would be necessary for reliable assessment. Teenagers who felt insecure about their competence in Fiji Hindi would presumably have been particularly reluctant to attend a social gathering at which they were expected to speak the language, and yet for the purposes of the research it was important that their proficiency should be assessed.

It was decided instead that the participants would be tested individually at home. One reason was that the first phase of the research, the self-report interviews, had been conducted successfully in the subjects' homes. This meant that their cooperation could be more easily enlisted by presenting the performance assessment — legitimately — as a follow-up to the earlier research. Secondly, the home was the domain in which Fiji Hindi was most likely to be used comfortably and so it offered the best opportunity to elicit fluent natural speech from the teenagers. These advantages contributed to the validity of the performance test, but on the other hand there was a trade-off in terms of reliability. The individual homes differed as settings for the test: other members of the family were usually present, various activities went on while the test was being administered, and interruptions could and did

occur. During several of the interviews music was playing in the background, telephone calls were received and visitors arrived. In one case, a very noisy bird chirped insistently. Thus it was difficult to achieve any consistency in the testing environment from one home to another and, when it came to rating the tapes, some of the teenagers' speech was not very audible.

Another potential threat to reliability was that the testing continued for an extended period of time. The interviewer made appointments to visit each home at a time that would suit the subject and the family. Although this produced a high participation rate, it also meant that it took about twelve weeks for the thirty-five interviews to be completed, since only one was conducted on any one day. As a way of addressing the problem of maintaining consistency over time, the researcher re-rated the interviews on tape during a concentrated period after they had all been completed, while the second rater assessed all the interviews on tape as well. Despite the extended testing period, there was not the concern that there would have been with a high-stakes test about maintaining the security of the test content. There was no real incentive for those who were tested later to try to improve their performance by seeking information about the test from earlier test-takers.

In the final analysis, then, although the home was an unsatisfactory setting in some ways from the viewpoint of reliability, these concerns were outweighed by the arguments for the validity and practicality of conducting the test in the teenagers' homes.

The interviewer

The test was administered to all of the subjects by the first author. This provided a greater degree of consistency than if two or more interviewers had been involved. She was obviously well-versed in the test procedure, having developed the test, carried out a pilot study with six subjects and revised the procedure on the basis of the experience gained. On the other hand, it is a well-established finding in social science research that a respondent's reaction to the personal characteristics of the interviewer will affect the responses given, and the same influences have been shown to operate in oral proficiency interviews (see, e.g., Scott 1986, Porter 1991). Thus, it is important to consider how the identity of this particular interviewer might have affected the results of the present study.

One great advantage she had was her insider status in the Indo-Fijian community, which meant that she was aware of the cultural and religious context of each home in which the interviews were conducted. She herself had immigrated to New Zealand after the 1987 coups in Fiji and therefore had shared a common experience with the majority of the subjects, which helped to establish rapport with them and their families.

In discussing the closeness of the relationship between the researcher and researched in the context of educational ethnography, Burgess (1989:60, citing Punch 1986) points out that access to and acceptance by members of a community are pivotal. This is especially true when community links are as strong as they were in the Indo-Fijian case. As an insider, the interviewer found it relatively easy to make contact with prospective subjects, who often knew a member of her family. She had additional legitimacy because contacts were made through respected community leaders. Her gender and religious background were also significant. Since Indo-Fijian parents are fiercely protective of their daughters, access would have been much harder to obtain, particularly with young Muslim females, had the interviewer been male and non-Muslim.

A disadvantage of being an insider, however, was that the interviewer knew some of the subjects personally. In these cases, care was necessary to prevent the interview from becoming too informal and to avoid the tendency to predict responses from her own knowledge of the teenagers and their families. In one case where the interviewer knew the family well, the interview was perceived as a social occasion, to the point where other family members were eventually asked to leave the room as they were affecting the course of the interview and the quality of the responses that were being elicited from the teenager.

Another potential drawback of the interviewer's identity was a sociolinguistic one. Among younger and better educated Indo-Fijians, both in Fiji and Wellington, Fiji Hindi/English code-mixing is the norm; young Indo-Fijians use English words and phrases in their Fiji Hindi speech among themselves all the time (Shameem 1994:402). If the interviewer had been an older person with a traditional demeanour and an apparently limited command of English, it would have been reasonable to expect that the teenagers would use relatively "pure" Fiji Hindi, at least to the extent of their competence in the language. However, given that

the interviewer was actually in her 30s, was modern in her ways and obviously highly proficient in English, code-mixing came quite naturally to her and to the teenagers in the interview situation. It would have been inauthentic to have imposed on them a requirement that only Fiji Hindi should be used. Thus the interviewer allowed the conversation to flow, without any requests or reminders to use Fiji Hindi and without a conscious effort on her part to avoid using English.

In assessing the teenagers' performance, though, the two raters had to be careful to distinguish between two ways of interpreting the occurrence of English lexical items in the teenagers' speech. One was that, as just noted, it represented the normal speech behaviour of fluent bilingual members of the community. The other possibility, of course, was that the teenager was using English words to compensate for a lack of knowledge of the equivalent terms in Fiji Hindi. In order to help distinguish these two possibilities, a vocabulary translation task was included as part of the test. There were ten words to be translated into English and another ten to be rendered in Fiji Hindi. Both sets were thematically related to earlier parts of the test and consisted of lexical items for which English had been used by the subjects who participated in the pilot study. The vocabulary task therefore indicated whether the Fiji Hindi term was known by the teenager, even if it had not been used in the context of the test.

There was a broader issue here that was relevant to the researcher in both of her roles, as interviewer and as rater: how to define native proficiency in Fiji Hindi, as represented operationally by the highest level on the rating scales. Both the researcher and the second rater belong to a generation of young adult Indo-Fijians, aged roughly between 20 and 40, for whom Fiji Hindi is their mother tongue but who have been largely educated through the medium of English. Their high degree of bilingualism has inevitably affected the quality of their Fiji Hindi in various ways, when compared with the speech of their parents and grandparents. The older generation, especially those living in rural areas of Fiji and with a limited knowledge of English, speak varieties of the language that are more conservative — and "purer" — than those of younger people. In one sense, command of these conservative varieties comprises the highest existing level of proficiency in Fiji Hindi. However, while younger adults can generally understand these varieties (as

indicated by their performance in the structured listening task), they do not speak that way themselves and it seemed quite unrealistic to expect that teenagers living in Wellington could attain that kind of proficiency because they have little if any exposure to the speech of the older generation in their new environment. Thus, for the purposes of the assessment, the highest attainable proficiency level was defined in terms of the competence in Fiji Hindi of the younger adult generation, to which the researcher and the other rater belonged.

The mechanics of the assessment

There are a couple of technical aspects of the assessment process that need to be discussed. The first is a linguistic one, deriving from the non-standard preliterate nature of Fiji Hindi. In the first phase of the research, while the interview schedule was being developed, an attempt was made to translate the questions into written Fiji Hindi. This exercise was eventually abandoned because of the difficulties created by the lack of an established script for the language and the lack of translation equivalents for many of the terms used in the questions. Thus, the interview schedule was written only in English. When the teenagers were being interviewed for the first phase of the research, they were given a written copy of the questions, since it seemed reasonable to assume that they were all proficient readers of English — and in fact they generally preferred that the interviews be conducted in English as well. In the interviews with their mothers, whose proficiency in English was more variable, there was more Fiji Hindi spoken, and the interviewer rehearsed in advance oral translations of the questions that mothers with limited English were expected to find difficult to understand.

Similarly, for the performance test, the assessment materials used by the interviewer and second rater — the rating scales, scoring sheets and assessment guidelines — were written in English. On the other hand, the instructions and the questions for the structured speaking and listening tasks needed to be presented orally to the test-takers in Fiji Hindi. In this case, the solution was to write the instructions and questions in an informal "semi-script" using Roman letters (which was the same strategy adopted by Raymond Pillai for his play *Adhura Sapna*). During the test, the interviewer read aloud the items in Fiji Hindi as written in this form.

These, then, were the pragmatic solutions to the problems posed by the non-literate nature of the language being assessed.

The second technical aspect of the assessment was the making of an audio recording of the interview. This is normal practice in oral proficiency interviews today, in order to allow for later re-rating of the test-taker's performance by the interviewer and/or other raters. It could not necessarily be assumed that the test-takers in this study would be willing to allow themselves to be recorded in the interview. However, in keeping with the standard requirements for research involving human subjects, the interviewer informed the test-takers beforehand of her intention to use the tape recorder, with an undertaking that the recording would be kept confidential, and they all gave their consent. Everyone had had the experience of being recorded in the first phase of the research, and any wariness they exhibited about the taping of the performance test tended to disappear when the interviewer maintained an informal style of speech after the recorder was turned on. In a post-test evaluation, only one person expressed any concern about the presence of the recorder.

Another issue that arises in relation to the audiotaping of performance tests is the comparability of live and tape-based assessments. There is evidence, for example, that raters assess taped performances more harshly than live ones (see, e.g., McNamara 1987:100-103). In rating audiotapes, one is forced to focus on the test-taker's speech without having access to the facial expressions, body language and other non-verbal aspects of communication that would have accompanied the speech in the live situation. This tends to make the rater more aware of structural errors in the test-taker's speech.

For this test in particular, there were several factors that contributed to the difference between the two conditions from the raters' viewpoint. A practical difficulty with rating a number of the tapes was, as noted above, the amount of background noise created by other members of the household, which was more distracting on the page than during the actual interview. Secondly, the reviewing of the tapes after all the interviews were completed allowed the interviewer to assess them within a much shorter time period than was possible in the live condition. This offered the opportunity to apply more consistent standards to the rating of the tapes.

Another advantage for the interviewer in listening to the tapes was that she was able to gauge the amount of English/Fiji Hindi code-mixing

and switching that had occurred more accurately than during the interview itself. In speech communities where code-switching is the norm, fluent bilinguals typically do not switch from one language to the other at a *conscious* level. As Heller observes, "many people who codeswitch are not aware of their behaviour until it is brought to their attention - and even when it is they often resist the notion that they really speak like that" (1988:6). In addition, those who have a general awareness of the phenomenon may be quite surprised to find how much of one language they used in a particular conversation that they had perceived to be conducted basically in the other one. This was the experience of the interviewer in the present study and, in re-rating the tapes, she sought to establish whether particular switches reflected fluent bilingual usage or some lack of proficiency in Fiji Hindi on the part of the test-taker.

In the light of these various considerations, it is interesting to note that the tape-based ratings were in fact lower on average than the live ones. The means and standard deviations for the listening and speaking components respectively are set out in Tables 1 and 2.

	Interviewer's live rating	Interviewer's tape rating	2nd rater's tape rating
Mean	5.11	4.83	4.94
SD	0.96	0.98	0.83

Table 1: Summary statistics of the three ratings of the Fiji Hindi listening test (N=35)

	Interviewer's live rating	Interviewer's tape rating	2nd rater's tape rating
Mean	4.43	3.97	4.00
SD	1.40	1.56	1.28

Table 2: Summary statistics of the three ratings of the Fiji Hindi speaking test (N=35)

The second rater's ratings were close to the interviewer's tape-based ones, although they had a somewhat smaller variance. The intercorrelations among the various ratings were all at 0.90 or above, indicating that overall there was a level of both intra-rater and inter-rater consistency in the assessments that was quite impressive, given the lack of previous experience of assessing proficiency in Fiji Hindi.

FEEDBACK ON THE TEST

The test-takers

At the end of the test, the teenagers were given an evaluation form in English containing nine open-ended questions to elicit their reactions. (A copy of the form is in the Appendix.) Most of them reported feeling comfortable during the test. They generally appreciated the way that the interviewer shared personal information about herself with them and in that way modelled the kind of responses that they were expected to give. They described her as a friend, a big sister, a person whom they felt they knew well and one whom they could talk to without restraint. Only a small number said that they felt restricted or pressured as a result of their limited knowledge of Fiji Hindi. And, as previously noted, there was just one person who expressed any concern about the tape recording.

With respect to the task of describing the making of an omelette, a large number of respondents reported that the pictures were clear and easy to follow. However, in the actual administration of the task, there turned out to be a problem with vocabulary. The pictures depicted a number of objects, such as a frying pan, butter, a whisk and salt and pepper shakers, which are not commonly found in Indo-Fijian kitchens. Even fluent speakers found it difficult to supply Fiji Hindi words for these items. Thus, even though the activity in the pictures was culturally appropriate, the way that it was portrayed created unforeseen lexical difficulties which reduced its value as a stimulus for Fiji Hindi speech production.

In the story retelling task, the difficulties were more cognitive than language-based. About two-thirds of the test-takers commented that the short story was written in simple clear English. However, more than a third of them had problems with summarizing the ideas in the story and/or translating them into Fiji Hindi. Part of the reason seemed to be that

c of high proficiency that he mentioned was the ability to respond to and express the unique sense of humour associated with the language, as well as its idioms and nuances.

CONCLUSION

In conclusion, it is useful to review what has been discussed in this paper in terms of reliability and validity. There were several potential threats to the reliability of the test. One was that the test was not conducted for any educational purpose, which meant that the test-takers had no instrumental reason for performing to the best of their ability. Secondly, the interviews were each conducted in a different location over a considerable period of time. The setting produced various distractions and interruptions in numerous cases, affecting both the interview itself and the quality of the tape recording for later rating. In addition, the conversational part of the interview was relatively unstructured and the test-takers were encouraged to take the initiative in introducing topics that were different from the ones that had been incorporated into the design of the test. Although there was a small pilot study before the research interviews began, the interviewer was in the position of training herself as she went along to administer the test and carry out the live assessment in the most effective manner.

Set against these limitations were other factors that contributed to reliable assessment. The researcher conducted all the interviews herself, and she had carefully developed the test procedure and the rating scales in conjunction with each other. While the test-takers may not have been aware of the structured nature of the conversation, the interviewer guided that part of the test through a series of topics linked to ascending levels of the proficiency scales. At forty minutes or more, it was quite a long test, providing a large sample of performance on which to assess the test-taker's oral proficiency. Furthermore, the audiotaping allowed for both intra-rater and inter-rater checks on the reliability of the assessments; these both revealed a high degree of consistency in the ratings that were given.

Turning to validity, there were some interesting conceptual issues to resolve in implementing this test. One was to define what represented

the highest level of proficiency in Fiji Hindi that could reasonably be achieved by the test-takers, given their age, their bilingualism and their residence in New Zealand rather than Fiji. Another was to consider how the phenomenon of code-mixing should be dealt with, both in administering the interview and in assessing the test-takers' performance. In a normal language test, a proficient speaker is expected to be able to speak consistently in the target language, whereas in this case fluent mixing of English lexical items in Fiji Hindi speech could be interpreted as expressing a high level of proficiency in Fiji Hindi as well as in English. It would have been sociolinguistically invalid to have attempted to elicit "pure" Fiji Hindi from the test-takers. A third issue, which relates to content validity, was the selection of culturally appropriate topics and tasks for the test. The length of the test allowed for a representative range of content to be included; however, experience with the test indicated that naturalistic conversation worked more effectively as a stimulus for authentic speech in Fiji Hindi than the structured speaking tasks. One further question was where best to administer the test. The testing of the teenagers in their own homes created some reliability problems, as noted above, but on the other hand it was the most suitable setting from a sociolinguistic viewpoint and the place where they were most likely to feel comfortable speaking their mother tongue.

There is much more that could be said about validity in relation to the Fiji Hindi test but that would take us well beyond the scope of this paper. As explained earlier, the test was part of a larger exercise to validate the subjects' self-reported ratings of their proficiency in the language. The comparison of the self-report ratings with those from the test showed a high level of consistency in the case of listening. For speaking there was a greater discrepancy, caused largely by the fact that numerous teenagers who had rated themselves at level 6 were assessed in the test as being at level 5. This difference in judgement may have resulted from the previously discussed problem of defining the top level of proficiency in Fiji Hindi.

It can be concluded overall that this performance test proved to be an effective instrument for its intended purpose and it provides a model for the development and administration of tests for other vernacular languages that lack a tradition of formal assessment.

REFERENCES

Bachman, L.F. (1990) *Fundamental considerations in language testing*. Oxford, Oxford University Press.

Burgess, R.G. (1989) Grey areas: ethical dilemmas in educational ethnography. In R.G. Burgess (ed.) *The ethics of educational research*. London, The Falmer Press.

Fasold, R. (1984) *The sociolinguistics of society*. Oxford, Basil Blackwell.

Hakuta, K. and D. D'Andrea (1992) Some properties of bilingual maintenance and loss in Mexican background high-school students. *Applied Linguistics* 13,1:72-99.

Heller, M. (1988) Introduction. In M. Heller (ed.) *Codeswitching: anthropological and sociolinguistic perspectives*. Berlin, Mouton de Gruyter.

Holmes, J., M. Roberts, M. Verivaki and 'A. 'Aipolo (1993) Language maintenance and shift in three New Zealand speech communities. *Applied Linguistics* 14,1:1-24.

Kloss, H. (1968) Some notes concerning a language nation typology. In J.A. Fishman (ed.) *Readings in the sociology of language*. The Hague, Mouton.

McNamara, T.F. (1987) Assessing the language proficiency of health professionals: Recommendations for reform of the Occupational English Test. A report submitted to the Council on Overseas Professional Qualifications. Melbourne, University of Melbourne, Department of Russian and Language Studies.

Porter, D. (1991) Affective factors in the assessment of oral interaction: gender and status. In S. Anivan (ed.) *Current developments in language testing*. Singapore, SEAMEO Regional Language Centre.

Romaine, S. (1989) *Bilingualism*. Oxford, Basil Blackwell.

Scott, M.L. (1986) Student affective reactions to oral language tests. *Language Testing* 3,1:99-118.

Shameem, N. (1992) *Arrival and other stories*. Wellington, Wellington Multicultural Educational Resource Centre.

Shameem, N. (1994) The Wellington Indo-Fijians: language shift among teenage new immigrants. *Journal of Multilingual and Multicultural Development* 15,5:399-418.

Shameem, N. (1995) Hamai log ke boli. Our language. Language shift in an immigrant community: the Wellington Indo-Fijians. Unpublished Ph.D thesis. Victoria University of Wellington.

Shohamy, E. and O. Inbar (1991) Validation of listening comprehension tests: the effect of text and question type. *Language Testing* 8,1:23-40.

APPENDIX

I. The Stimulus Text for the Story Retelling Task

Arrival

The girl looked out eagerly as the plane circled over Auckland Airport. She had very little idea of what to expect. She had not thought about it much up till now. Her mind had been too full of thoughts of home and family and leaving so much of her life behind.

As she peered out of the little window beside her, she noticed the grass was a bright green -almost yellow. Tiny white dots stood out against the green.

"They must be the sheep," she thought. She remembered Bhaini telling her that there were a lot more sheep than people in New Zealand. She had never seen sheep before.

"So different from home," she thought. Her last glimpse of Fiji had been of the cane fields and of the muddy Nadi River, snake-like in the distance.

Here even the grass was a strange shade of green. And there, further away, were a lot of houses and buildings - many more buildings than Suva had. Taller too, many of them.

The plane was coming in to land. Her heart was thump-thumping. Her chest felt sore. She couldn't breathe properly. She felt very frightened.

"Please let Bhaini be waiting for me," she prayed silently to herself. This was the first time she had ever been out of Fiji. Her only flying experience before this had been on a very short Fiji Air flight to Labasa, many years ago, when her mother had taken her to Nana's funeral.

The plane came to a stop. She clutched her pale blue Fijian passport and her customs and immigration cards nervously. The captain said no-one could leave the plane until it was sprayed. Her chest was feeling worse. The pain moved down to her stomach. She felt sick.

Two men in blue and white uniforms came into the plane. They walked down the aisle spraying from spray cans in both directions. The elderly woman

next to her seemed to have difficulty breathing. She had placed the achra of her sari over her mouth.

"Are you alright?" asked the girl, in Hindustani.

"I have asthma," the woman replied, her voice muffled behind her achra.

The girl knew what that felt like. Her mother also had asthma. Sprays always affected her too.

The men stopped spraying. The woman next to her seemed better. The girl helped her to collect her bags. She had to stand up and reach into the overhead locker for a larger bag.

The steward opened the door. It was time to go. In her concern for the woman she had forgotten to feel scared. Now the fears returned in a rush.

This was a new country - one that she had only heard about from her older sister who had migrated here three years ago. Bhaini had sent home some photographs and picture postcards of mudpools and geysers. But that was all she knew.

She picked up her new handbag - a parting gift from her younger sister - and hoisted it over her right shoulder. She walked with a heavy heart towards the open door. She hardly noticed the steward at the door nodding and smiling good-bye to her.

(Shameem, 1992)

II. The Feedback Form for the Test-takers

TEST EVALUATION FORM: RESPONDENT

Part A: Conversation

1. Did the session feel comfortable to you? Did you feel pressured at any point? When?

2. What made it comfortable/uncomfortable for you? How could this be changed?

3. Were the statements made by the interviewer about herself helpful in helping you feel more comfortable?..... in helping you to think about your own answers?.....

4. When you were asked to retell the story 'Arrival', were you clear about what you had to do?

5. Was the story simple enough to read and retell? What do you think made this possible?

6. When you were asked to describe the pictures on making an omelette, what if anything, did you find most difficult ?

Part B: Listening Comprehension

1. Was the recording clear and easy to follow?

2. Did being told the questions before you listened the second time help you in answering the questions?

3. Were the section divisions short enough?

III. Fiji Hindi Test Descriptors

Since proficiency is essentially to be assessed on a continuous scale, the symbol + may be used at any level to indicate that proficiency is judged to exceed the specifications for that level, but falls short of the requirements of the higher level.

Listening

0 No understanding of Fiji Hindi. Unable to demonstrate comprehension at even the most basic level.

1 Demonstrates understanding of basic courtesy requirements, for example greetings and 'yes' and 'no' in FH. Has difficulty understanding utterances which are longer. Shows this by lengthy pauses or frequently asking for clarification and/or repetition. May ask for translation to English. Has difficulty understanding play extracts even at the trivial level.

2 Understands questions/statements/discussion on self and family when interlocutor speaks very slowly. Can understand short, simple statements about the interlocutor, her family and her background. Understands appropriate FH sounds for request for further information, 'yes' and 'no'. Demonstrates understanding of question tag, na?, hana? and single digit numbers, one to ten. Understands play extracts from Adhura Sapna only at the trivial level.

3 Minimum social proficiency. Is able to understand a broader conversation on immediately familiar places and things. Understands main ideas but frequently asks for repetition of longer questions and more complex statements. Experiences

difficulty understanding probe questions and descriptive statements - frequently seeks clarification. Understands time phrases, yesterday, today, tomorrow. Recognises base colours, black, white, red. Understands questions requiring local answers in FH play, but experiences some difficulty in understanding global questions.

4 -Basic social proficiency. Is able to understand a native speaker in a face to face situation, speaking at a normal rate. Asks for some repetition and clarification but understanding generally appears sound. Understands requirements in probe questions. Understands statements and questions involving broader but still familiar topics such as descriptions of place of birth, Wellington, school. Able to understand questions dealing with well travelled distances and size of a familiar place such as school. Understands questions requiring global answers in FH play. Experiences some difficulty with understanding third person speech at normal utterance rates - asks for repetition of sections of play.

5 Social proficiency. Has no difficulty with understanding interlocutor's contribution to the conversation. Understanding is easy, spontaneous and sound, although inadequate FH vocabulary base may create unease and quick word substitution. Paraphrase or translation may be needed. May show some evidence of breakdown in comprehension of third person FH speech in Adhura Sapna. Occasionally baffled by colloquialisms, more difficult/less used/rural variety/older variety of FH words and phrases (LCT, Q 1, 5, 10, 11, 13, 15, 16, 17, 18). Comfortably understands questions requiring global answers (making inferences) in FH play.

6 Native social proficiency. Is able to understand all FH speech at normal to fast utterance rates on a broader range of appropriate topics, including the issue of FH maintenance in NZ. Has the broad vocabulary range to understand FH utterance without needing repetition, translation or paraphrase. Likely to know day of the week, time of day, other colours and double digit numbers.

Comprehends FH humour (LCT, Q 11-G, 15-G), rural/urban varieties (Q 13-G, 18-G), colloquialisms (Q 10-G, 11-G) idioms (Q 15-G). Can identify mood/tone of discourse (Q 8-G, 9-G, 12-G). Comprehends subtlety of meaning (Q 1-G, 5-G, 6, 20-G), cultural references (Q 1-G, 7-G, 8-G, 10-G). Able to discern and comprehend older FH speech variety, including lexis, idioms (Q 15, 16, 17). Sensitive to implications of variety used - level of education, age (Q 17-G), setting (Q 18-G).

Speaking

0 No speaking ability. Unable to respond to greetings or simple questions requiring 'yes', 'no' answers.

1 Responds appropriately to FH greeting. Is able to reply to simple questions requiring 'yes', 'no' answers.

2 Responds to questions on self and family in FH with little hesitation. Is able to identify immediate family relationships. But minimal response - tendency to respond in single words. Vocabulary used is appropriate to the immediate context and topic, although pronunciation of borrowed/transferred English words has NZ accent. Expression lacks autonomy - largely reliant on words, expressions and structures used by the interlocutor. Frequent long pauses. Appropriate response to question tag, hana...na..? Appropriate response to non-verbal communication. Knows and is able to use numbers 1 - 10. Uses the culturally appropriate form aap and ji or tum and haa in response to the terms used in the interview.

3 Minimum social proficiency. Participates in basic social intercourse on familiar places and things. Still expecting simple sentence answers but evidence of wider vocabulary - related to familiar social/cultural issues. Able to use time phrases kal, aaj, bhiyan. Responds with appropriate vocabulary. Knows base colours karia, uJar, lal. Pronunciation of borrowed/transferred English words has NZ accent. Seeks clarification appropriately. Has difficulty in responding to probe questions - hesitates with

wide pauses; or repeats interlocutor's statements. Shows some ability to re-tell the short story Arrival and describe parts of the process of making an omelette. Responds appropriately to questions requiring trivial answers in FH play extracts from Adhura Sapna (LCT Q. 2, 3, 14). Basic local responses in FH to questions in play extracts.

4 Basic social proficiency. Still on familiar topics but is able, with frequent prompting, to make a greater contribution to the conversation. Responds appropriately to probe questions with some hesitation and circumlocution - but is able to get message across. Frequent use of English conjunctions in FH sentences. Is able to identify in FH, gist of the short story Arrival - prompting may be necessary. Is able to identify in FH the process of making an omelette but has some difficulty with describing the procedure - frequent hesitation and/or a high input of English words. Makes local responses to questions on FH play extracts with ease and insight. Basic FH responses to global questions on FH play.

5 Social proficiency. Is able to make comfortable contribution to native speaker conversation, although may need some prompting. Is able to self-correct performance errors (errors which occur naturally in conversation rather than those caused by lack of competence). Words transferred from English are acceptable in the context they are used and have the appropriate FH accent. Tendency to use the English conjunctions 'but' and 'because' in FH sentences. Is able to retell a simple English short story Arrival. Is able to use it to make some reference and/or comparisons to own life. Has difficulty using sequence words (e.g. pehle, tab, fir) or finding appropriate words to describe the process diagram. Good global responses to questions in FH play.

6 Native social proficiency. Able to carry out a conversation with a competent native speaker without hesitation or unnecessary circumlocutions. Words transferred from English, or English and FH words which may be placed in a relationship together

(collocations), are used with the appropriate FH accent and stress patterns. Stress and intonation patterns conform to native speaker standards - whether urban or rural variety. Is able to tell the time, identify the day of the week and knows easy double digit numbers (10-20) in FH. Comfortably summarizes and translates plot of English short story Arrival and is able to relate short story to own life and draw parallels or compare. Able to describe a process diagram with high FH lexical (word) accuracy and appropriate sequencing. Is, with fluency, able to discuss and defend opinion on the pertinent issue of FH maintenance among younger users in Wellington. Easily draws inferences and makes appropriate global responses to questions on FH play. Easily answers general summative questions on FH play. Able to use appropriate idioms in response to native speaker questions (Q 7-G, 20-G).

Respondents will be asked to identify the day of the week, the time of day, their age and colours other than black, white and red. Their ability to respond to these queries in FH will be placed at level 6. However if they have shown competence at level 6 in most of the other tasks but are unable to respond to these questions in FH, they will not be penalized since the use of these features in current FH, particularly in urban areas, tends to be in English.

SPECIFIC-PURPOSE LANGUAGE PERFORMANCE TESTS: TASK AND INTERACTION[1]

Tom Lumley and Annie Brown
The University of Melbourne

ABSTRACT

Oral performance tests designed to assess candidates' ability to function in a particular professional role typically involve role plays which simulate aspects of professional interaction between a candidate, in his/her professional role, and an interlocutor, in the role of client or consumer of the professional services. The validity of such tests is dependent both on the incorporation of appropriate content and on the extent to which the interaction between candidate and interlocutor simulates authentic professional interaction. Materials for such role plays are commonly produced according to specifications developed by language testers following consultation with professional informants, but there has been little investigation of how well tasks reflect their purpose (i.e. the aims/intentions of the professional experts). Test interlocutors, generally language teachers, normally undergo some form of training in the test administration procedure, but there has been little analysis of how well their behaviour reflects client behaviour.

This paper reports on an investigation into the role which expert professional informants can play in providing post hoc content validation of tasks used in this kind of occupation-specific test, focusing on both the content of the tasks and the behaviour of the interlocutors. The study draws on data from the oral interaction component of an occupation-specific test, the Occupational English Test (McNamara 1996), a 4-skills test for overseas-trained health professionals (with attention in this study to nurses). The study involved a two-day workshop where representatives of the nursing profession, in conjunction with the researchers, analysed both written test materials and audiotaped test data, contributing to an analysis of the realism of the tasks and the features of interlocutor behaviour.

INTRODUCTION

Performance testing of oral language skills has become increasingly commonplace within the context of occupational language proficiency assessment for certification (see, for example, Elder, 1994; McNamara, 1990, 1996; Brown, 1994; Wesche, 1987, 1992; Douglas forthcoming).

ARAL Series S Nº 13 (1996) 105-136

Consistent with current views of language which emphasise the importance of communicative skills in addition to the mastery of the linguistic system, performance tests are claimed to enable candidates to demonstrate what they 'can do in their second language, rather than what they know about it' (Wesche, 1992:104).

Where the context of future use is relatively definable and describable, the rationale behind performance tests is that predictive validity is likely to be higher where the candidates are faced with a simulation of relevant communication challenges of the workplace. Hence they required not only to demonstrate knowledge of the linguistic system, but also to demonstrate their ability to use language appropriately and skilfully in relation to specified communication contexts. In other words, simulating target performance will result in relevant communication skills being tapped.

The pursuit of authenticity is, then, a consequence of attempting to capture within a test the complexity of the integrated communication skills required in the target situation. However, writers on performance testing agree that authenticity is an unobtainable goal, and as a result, 'authentic' or 'real-life' language testing has come under attack from various writers for failing to make the necessary distinction between the test and the criterion, between behaviour and ability (Bachman 1990:308). However, as Davies argues,

> 'There is no need to fall into the Bachman trap of confusion between ability and behaviour. Work sample tests meet the basic language testing requirements: they test language ability, in context and with adequate coverage. While avoiding 'real life', such tests have claims as performance tests. The problem for test construction is therefore practical rather than theoretical, a question of appropriate task sampling.'
>
> (Davies 1995:15)

He concludes that

> 'The art of language testing comes from better descriptions of language features *and* from more valid sampling of critical contexts of language in use.'
>
> (Davies 1995:18).

The problem remains, though, that while simulation of the target context is necessary if we are to infer something about candidates' ability

to interact in specific communicative contexts for specific purposes, the fact that authenticity is ultimately unachievable begs the question of *the extent to which* we can use test performance to make inferences about performance in the target context. To answer this question requires detailed comparisons of test and real-life performances from a variety of perspectives.

This paper attempts to shed some light on the relationship between test and real-life performance through an investigation of two threats to authenticity, namely:

a) the extent to which target communication contexts can be simulated within the constraints of the testing situation, and
b) the extent to which the interaction between candidate and interlocutor reflects real-life interaction.

A recurrent problem in the testing of occupational language skills is that of the incorporation of occupational or subject matter knowledge into test materials. Since the language test writer (typically a language teacher or testing specialist with a language background) seldom has much expertise in or familiarity with the occupational field, specialists from the relevant occupational field are often consulted during test development. Roleplay materials may be developed by, or in conjunction with, occupational informants, or may be subjected to a process of vetting by such informants prior to trialling. This process of materials development is by no means straightforward as the testing situation imposes constraints on the interaction, one of these being the amount of contextual information that the participants can be expected to deal with in an oral testing context. While the focus in a language test is not usually on the display of occupational knowledge, a certain amount of contextual background to the communication task is a prerequisite to the development of the test interaction. To be incorporated successfully into the roleplay materials used by the two participants (candidate and interlocutor) this information needs to be concise, accurate and relevant. Particularly where a range of alternative roleplays are to be used, fairness to candidates (which relates to both the validity of a given test as an index of ability and to the whole testing process in so far as it reflects or contributes to social equity) requires that all materials be equally accessible and feasible as stimuli.

A further aspect of the validity of the roleplay as a legitimate representation of occupational interaction is the extent to which the interviewers' linguistic behaviour is consistent with the types of linguistic behaviour the candidate is likely to encounter in the target situation. To this extent, the interviewer constitutes one facet of the test method (Bachman 1990). We would argue, on the basis of our own experiences as test developers and test interlocutors, that the interlocutors may pose a threat to the validity of the test, given that they are generally drawn from the language teaching profession: the perspective this expertise gives them (a 'teacherly' view of second language learners, and all that this might imply), together with lack of experience in the occupational situations simulated in the test, may result in the interaction being somewhat different from that intended by the test developers and writers. Brown (1995) reports that as *raters*, language teachers have a different perspective and award grades somewhat differently from raters drawn from the occupation for which the test is designed (in this case a test for intending tour guides). It is, therefore, quite conceivable that as *interlocutors* they may differ in the way in which they interact with the candidates.

While we are not aware of any studies focussing specifically on the extent to which test interlocutors in occupational LSP tests adopt the role prescribed and establish a suitable interaction, a number of studies have investigated interlocutor variability in general. Ross and Berwick (1992) and Ross (1992), for example, have discussed the extent to which test interlocutors accommodate their language output to the candidate's perceived level of ability, and the implications of this for test validity. Young and Milanovic (1992) have investigated interlocutor differences in interactional contingency, goal orientation and dominance. Caffarella (1994), Filipi (1994), Lazaraton & Saville (1994) and Morton, Wigglesworth and Williams (forthcoming) have all observed significant differences in interviewer language and behaviour and found this to influence the language output of candidates.

These two issues, that of the extent to which the role play materials accurately represent authentic or believable scenarios, and that of fairness in terms of challenge presented by the interlocutor and his/her linguistic behaviour, both reflect aspects of the validity of a test. As Cumming and Berwick point out,

> 'various practices, modes of argumentation, sources of information and complex conceptual issues exist for establishing the validity of language testing instruments, procedures and their uses'.
>
> (Cumming and Berwick 1996: 1)

This study expands the role of occupational experts past that of *a priori* validation and investigates the role they can play in *a posteriori* content validation through examination and evaluation of the quality of the materials and the characteristics of the interactions.

To summarise, the questions we asked in this study centred around two issues:

a) given the constraints in transferring occupational interaction to a testing situation:
   i) to what extent is it possible to simulate occupational communication tasks effectively, and
   ii) what problems are likely to arise in the development of the test materials?

b) i) in what ways do interlocutors react to the demands of the testing situation, and
   ii) what impact does this have on the interaction?

METHODOLOGY

The test

Our research in this area involved the oral component of an occupation-specific language test, the Occupational English Test (OET) for health professionals.

The OET (McNamara 1990, 1996) is used as part of the professional accreditation procedure for non-English-speaking background overseas-trained health professionals wishing to practise in Australia. The test is currently used by 11 health professions, of which medical practitioners form the largest group, and nurses the second largest. The test consists of four sub-tests, one each for reading, writing, speaking and listening. This study considers only the speaking sub-test used by nurses.

The speaking sub-test consists of a non-assessed warm-up phase (taking the form of a brief interview) and two roleplays (both assessed)

of about 5 minutes each, simulating professional consultations between the candidate, in her professional role, and an interlocutor, in the role of patient or relative of a patient (client). Each candidate is presented with two roleplays randomly selected from a pool of materials. The materials used in the nurses' speaking sub-test are prepared by practising nurses, in consultation with the test developers.

The test tasks concentrate on some aspect of management of a patient's health and the patient's or relative's concerns. The interaction between candidate and interlocutor involves the use of particular linguistic skills: candidates are required to deal with a difficult situation involving a point of tension and (consequently) a high level of emotional involvement on the part of both candidate and interlocutor. The objective is thus to push candidates into negotiation and to find out whether they have the linguistic and communicative resources to deal with such challenging situations (requiring the ability not only to make suggestions but to persuade, reassure, pacify, etc.).

For example, the interlocutor's roleplay instructions may include:

1) resistance to or nervousness about proposed management of a health problem (such as taking particular drugs);
2) concern or worry about a suspected or diagnosed health problem (a skin cancer, a dietary problem); or
3) unwillingness to accept a particular diagnosis (a heart attack, for example) or advice (eg on giving up smoking).

The interlocutor (patient/client/relative) may be instructed, then, to act worried, unconvinced or unco-operative. While this sort of situation is commonly perceived as 'challenging' for non-native speakers, the ability to deal with these kinds of interpersonal issues was identified as crucial by professional informants (McNamara 1990) who were closely involved in both test design and initial production of the materials used.

Both candidate and interlocutor are given roleplay instructions. Figure 1 shows a sample of the instructions given to a test candidate and interlocutor[2]. As can be seen, the interaction concentrates on some aspect of management of a patient's health and the tension arising from the patient's or relative's concerns, rather than on the history-taking or diagnosis. Necessary background/professional information is in fact

CANDIDATE'S CARD

SETTING	Hospital Emergency Department
NURSE	You are talking to the son/daughter of 74 year old female patient admitted half an hour ago following a CVA[3]. The medical registrar has not yet arrived to examine her. You are responsible for monitoring the patient's condition (BP, vital signs, etc). Her condition has not deteriorated since admission, but she is comatose.
TASK	• Explain to the son/daughter the reasons for the monitoring procedures. • Reassure him/her that no active medical intervention is possible/appropriate at this stage. • Sympathise with the relative's concern and admit your uncertainty about the prognosis.

INTERLOCUTOR'S CARD

SETTING	Hospital Emergency Department
PATIENT	Your 74 year old mother has just been admitted to Emergency after a severe stroke. She is unconscious and her breathing is laboured. You are afraid that she will die. The doctor has not yet seen her and the nursing staff have carried out checks on blood pressure and vital functions but seem unconcerned. You are increasingly agitated and feel that your mother is being neglected.
TASK	• Convey your anxiety to the nurse and demand to know why so little is being done. • Ask about appropriate procedures for your mother's condition and possible outcomes. • Be insistent.

Figure 1: Sample Roleplay Card

provided to both participants, as the purpose for the test is to assess the nurses' language skills only. In other words, professional knowledge is deemed irrelevant in this test (being examined by the relevant professional bodies at a later stage). All interlocutors go through some form of training, in which this focus is made clear to them, and candidates are reminded of this at the beginning of the test. The interlocutors are oriented during their training to the need to focus on the point of tension in the roleplay, as described above. It is clear that the candidate cannot predict from the contents of her own instructions how the interlocutor will behave.

Data collection

The informants for the study were ten nurses, who provided written comments and judgements of test materials in a questionnaire, and then participated in a two-day workshop. All informants were very experienced in hospital practice and procedures and had all had considerable professional contact with NESB nurses.

The workshop conducted by the two researchers aimed to elicit detailed information from the ten nurses about their reactions to the test materials and interactions. The workshop took the form of a single group discussion. One of the researchers took principal responsibility for leading the discussion, while the other took notes on main issues of interest, recorded language assessments made by the nurses of the test performances presented in the workshop, and participated in the discussions as appropriate. Audio tape recordings were made of the workshop sessions. Because twelve people took part in the workshop, producing constant lively discussion, it was impossible to produce a reliable transcription of the sessions. However, the researchers were able, by reviewing the recordings subsequently, to identify issues of interest, to confirm or elaborate on points recorded in the notes written during the workshop, and, where possible, to provide examples of test interaction or informants' discourse to illustrate particular points.

The workshop contained the following stages:

1. Introductions, administrative procedures, participants' views of the context in which the test is used, purpose of workshop;

2. Detailed discussion of the written materials (role play cards), based on questionnaire responses provided by participants
3. Review of samples of test interaction, noting:
   a) the accuracy of participants in following the instructions provided;
   b) the realism of the interaction;
   c) the level of linguistic support given by the interlocutor;
   d) a proficiency rating for each candidate (not discussed in this study).
4. For each sample interaction, group discussion of notes taken by the researchers.

FEEDBACK ON THE WRITTEN MATERIALS

Before the workshop, participants reviewed four sets of roleplay materials (candidate and interlocutor instructions) and completed a questionnaire. At this stage they were asked to provide an evaluation of, and written comments on, the realism of the situation portrayed in each set of materials, and on the extent to which the instructions appeared to allow the interlocutor and the candidate to establish a believable role (as patient and nurse, respectively).

In the questionnaire, the occupational experts generally rated the written materials highly (mean ratings of between 4.7 and 5.8 on a six-point scale were provided for all questions), indicating that overall they considered them to be valid representations of professional situations. Typical comments about the materials included:

> 'I face this sort of situation every day in the ED (emergency department). Families are anxious and frightened and frequently confront nursing staff about what is going on.'

> 'This is a good scenario, gives enough detail of fact. Includes the more difficult emotional aspects.'

Information obtained from questionnaires in this way seemed to provide evidence of content validity.

In the workshop that followed, informants were asked to elaborate on their initial written comments, commenting on features of each role play that contributed to their perception of it as realistic or unrealistic, and making suggestions for improvements to the content; the resulting

group discussion revealed important qualifications to the general impression conveyed by the questionnaire responses, in that, as a result of the extensive perusal and discussion of each set of materials, problems came to light which had not previously been identified (or if so, only by one or two participants) in the content of some tasks.

Of the four roleplays examined in this study, two were identified as problematic. The problems were of two types: inaccuracies or inappropriate instructions in the description of typical procedures; and the amount of information given or the level of detail provided (in that either important points were missing or the situation was somewhat complex and could be misunderstood). These two roleplays will be discussed in turn.

One roleplay (see Figure 1 above) is set in the emergency department of a hospital, where the nurse's task is to calm and reassure the relative of an elderly patient who is unconscious and appears to have suffered a stroke. Three specific concerns were raised with this roleplay.

Firstly, the diagnosis of a stroke was unclear. Because the patient has not yet been formally examined by a doctor, the nurse is not authorised to discuss the diagnosis. However, the roleplay instructions include a requirement for her to make reference to the diagnosis. This instruction conflicts, therefore, with what might be considered good professional practice. Secondly, the nurse is instructed to reassure the relative with the advice that 'no active medical intervention is possible/ appropriate at this stage'. In reality, basic first aid measures would be taken (e.g. the use of a drip, dealing with laboured breathing). Thirdly, with regard to the instruction given to the nurse to provide reassurance, workshop participants referred to one strategy commonly used in reassuring patients and relatives, that of asking questions about the family and the patient's medical history. In the testing situation this approach is impractical: the interlocutor has no background knowledge about this fictitious patient and family on which to base answers; the need to keep the stimulus materials reasonably short makes it impossible to include enough information to enable the interlocutor to answer such questions; and the need for the candidate, not the interlocutor, to produce the major portion of the speech would render this undesirable. The advice the candidate is told to use, that 'no active medical intervention is possible/ appropriate at this stage', is unlikely to reassure any relative who is

concerned about the lack of action. In this sense, then, the test materials do not allow for 'realistic' interaction and the replication of a typical scenario; they are unable to simulate real life.

The implications for the candidate are that she might find it difficult to provide reassurance using the information provided, and thus to fulfill the task prescribed. An issue of fairness is created: the candidate may have difficulty deciding how to react to the instruction in this roleplay. Some candidates may be able to gloss over or ignore the inauthenticity; others, however, may be uncertain how to treat the situation, and as a consequence their performance may be affected. The identification of these problems in the workshop highlighted one difficulty in developing materials that are both realistic and brief for occupation-specific roleplays: the reduction of a complex scenario to a few lines of roleplay instructions can result in the failure to include necessary information. Similarly, the need to set up an interaction where the candidate is required to produce most of the language, plus the need to avoid overly detailed background information can render the roleplay atypical.

In the second roleplay the patient was a car crash victim who was instructed to behave uncooperatively, while part of the nurse's task was to talk to the patient about how the accident happened and to discuss his/her emotional state. The problem here was that although the counselling role prescribed for the candidate was described by some of the workshop participants as the 'textbook' response in this sort of situation, others pointed out that in their experience a nurse should deliberately avoid soliciting information about how the accident happened, on the grounds that such knowledge may influence her attitude towards the patient's well being. In other words, they felt that the instructions directed the candidate to behave unprofessionally. A further problem with this roleplay concerned the description of procedures already undertaken. In the roleplay instructions possible internal injuries were alluded to; the candidate would need more definite information about these, so as to be able to make more authoritative statements about treatment. Again, the problem of providing abbreviated descriptions of hospital scenarios contributed to some loss of authenticity.

Despite the problems mentioned, this second task was perceived as demonstrating a high level of authenticity, and capable of generating appropriate discourse. This view seemed to be based on a perception that

in this case in particular a candidate could simply ignore the explicit instruction in the roleplay that conflicted with good practice; alternatively, by ignoring good practice they were still following a counselling role (a kind of communicative situation nurses need to be competent in). Nevertheless, the issue of fairness to candidates arises again; how candidates deal with the decision of whether to or how exactly to follow the roleplay instructions and how their decision affects their confidence will vary from candidate to candidate. It is perhaps not surprising that test writers who are occupational experts but who have not themselves been in a testing situation such as this (or even in many cases not learned a foreign language) may fail to understand the effect of their materials on the generation of interaction in the testing situation, and the importance of explicit detail and instruction. This problem can be overcome, perhaps, by ensuring that the test materials are reviewed by both test developers and occupational experts together.

The fact that of the four roleplays reviewed, two were seen as unproblematic shows that it is possible to write materials that successfully facilitate simulations of professional situations. However, this is clearly not a straightforward activity, and not one that can be successfully conducted solely by either language testers or occupational experts. The study showed the value of the workshop in eliciting a thorough analysis of the materials. Even where the expert informants had provided written reactions to the materials in questionnaires, most had responded favourably, and with little comment or analysis. To describe potential problems such as those detailed above, and their possible consequences, would be quite time-consuming: perhaps materials are best reviewed not by questionnaire or by requesting written comment, which appears to lead to a fairly superficial consideration of content, but by interview or workshop. What is certainly clear is that expert opinion *is* required at both the needs analysis and drafting stage and also in subsequent validation of test materials. Analysis of the materials by workshop participants shou'd lead to revision of the content of roleplays in terms of the wording, the appropriateness and accuracy of the information/advice provided.

FEEDBACK ON THE TEST INTERACTION

The expert informants were asked to review 17 audio recordings of test interactions[4], in order to evaluate the extent to which the potential

problems identified earlier were realised. They were asked to comment on:

1. The accuracy of participants in following roleplay instructions
2. The realism of the interaction
3. The level of linguistic support provided by interlocutor.

Generally, they considered that both interlocutor and candidate followed the roles set out for them, and that the interaction in the test was representative of the range of possible kinds of interaction in hospital situations. However, in some of the interactions involving the first roleplay mentioned above, comments tended to confirm that the problems predicted for candidates were realised on occasion: the interaction at times became very repetitive, with the relative (interlocutor) demanding action but the candidate unable to provide either action or reassurance. The workshop participants generally felt that in this roleplay such behaviour was a result of the candidate being 'powerless', in the sense that she had no information (i.e., neither formal diagnosis, nor provisional diagnosis from a casualty officer, nor standard medical procedures) with which she could reassure the relative.

In one typical example[5] the candidate followed the instruction to sympathise with the relative's concern by repeating several times, with minimal variation in expression, that such a situation was to be expected. Later in the same interaction the candidate continually repeated that the doctor would come soon.

It is of course quite conceivable that, as a result of repetitions such as these, lacking in substantive content, an assessor might infer that the candidate had limited resources of grammar and expression. This inference might or might not be fair: it would be hard to say whether a candidate's language use truly represented her ability, or was limited as a result of the poor structure of this roleplay. The problem of non-suppliance and how to interpret it is a perennial one in language testing. It can, however, be minimised, by ensuring that the roleplay prompt contains adequate content.

In a second instance, forced to explain why the doctor had not yet arrived to provide a diagnosis, the candidate gave contradictory messages concerning what will happen to the patient:

line (Candidate 69-1: candidate's utterances)[6]

18 No, no . A nurse is, is watchin her . yeah . umm . umm . if doctor
19 come . he can't do anything at this moment because your mother
20 is unconscious . we just . um . to see how . is . she getting on
...
54 yeah . after doctor come , er he may be, she may be sent to . er .
55 intensive care unit.

The limited nature of the contextual information given in this roleplay may have forced this candidate to spontaneously invent information (not supplied in the roleplay instructions) in order to deal with a 'relative' (the interlocutor) who had embraced his role enthusiastically, and insistently demanded action. From the point of view of the occupational experts listening to the interaction, the realism of the action was affected by the candidate's inability to invent information which they found convincing. It remains unclear what effect this would have on assessors, who are instructed to consider neither the accuracy of the information nor the extent to which they find the candidate convincing. A fundamental question is raised here in relation to performance tests of languages for specific purpose; that of how possible it is to separate consideration of language ability and content knowledge. The issue arises not only in the construction of test materials, where it is necessary to include material that is both accurate and relevant, but in the rating process. It is impossible to say how assessors will be influenced by the content of the candidates' speech, but it is unlikely that there will be no influence (see Douglas and Selinker (1993) and Elder (1993) for further discussion of this issue).

It may be that language-trained assessors are less likely to be influenced by such inappropriate or inconsistent explanations as the one above than occupational experts, who are more likely to be influenced by content. If this is the case one could conclude that roleplays such as these may not be suitable in a 'strong' performance test of language, or one where the raters are drawn from the relevant occupational background.

In a third example involving this roleplay, the candidate also gave contradictory (and therefore unrealistic) information about the condition of the patient:

(Candidate 52-1)

N = Nurse (Candidate)
P = Patient (Interlocutor)

N: doctor explained that um (1.0)
(clears throat) this stages could er .. we could do only the monitoring but her deterioration and her .. proc- progressing so . hh . um .... that's all we can do
...
N: ... her general condition
P: ok . and how's that?
N: er she, she seems the same as what she had um . it's the same .umm . condition as . er a while ago I mean when she had a stroke . hh . but er .. little bit rittle bit of improvement I can see because her blood pressure is little bit lower down

It is unclear here whether the patient is deteriorating or improving. It is possible that this contradiction was the result of firstly, the candidate's desire to follow the roleplay instruction to reassure the relative, and secondly, the lack of specific information about the patient's condition.

The roleplay involving the crash victim, described above, also elicited comment about its realism in relation to the instruction requiring the interlocutor to be uncooperative. The nurses commented that in the reality of a busy hospital, uncooperative patients tend to be ignored. In the testing situation, neither refusal to communicate on the part of the patient, nor the nurse ignoring the patient, is appropriate. However, uncommunicative behaviour by a test interlocutor is likely to cause strain for the candidate, who would have to take more responsibility for initiating and maintaining the dialogue. For this to happen in some test interactions but not others would lead to unfairness in terms of the communicative demands placed upon candidates, yet the review of test interactions indicated that there were indeed substantial differences between interlocutors in terms of the degree to which they co-operated interactionally. This will be discussed in more detail later.

Nevertheless, in general the informants' view was that the variability in the behaviour of the interlocutors (in terms of co-operativeness) was all within the bounds of what might be encountered,

and as a result contributed to the validity of the test. They did, however, comment at times on the (in their view) inappropriate level of linguistic support provided to the candidate by a particular interlocutor. For example:

> 'she seems to be easy on the candidate.'
>
> 'the nurse was struggling along there, she (the interlocutor) was being too kind.'
>
> 'the daughter was nowhere near what she would be in reality.' (i.e., not as insistent and demanding as she would have been in real life).

In other words, while they saw variability as natural and therefore acceptable, they were nevertheless aware of the fact that such differences could result in different challenges for candidates, and at times even identified interlocutor behaviour as verging on unnatural where it appeared too supportive. They were certainly able to identify where the interlocutor did not follow the instructions, as in the third example above.

In the main, however, workshop participants' comments tended to focus on their assessments of the professional style or knowledge displayed by the candidate. This was clearly of greatest interest to them. One Indian candidate, for example, who spoke Indian English, elicited varied opinions from the occupational experts concerning her language proficiency. These opinions were mainly based, it turned out, not on her language proficiency but on the style she used in the test, which they labelled as patronising and completely inappropriate for a nurse.

Her behaviour was also seen as inappropriate because it was extremely blunt, questionably accurate, and certainly far from reassuring as is evidenced in the following extract where the candidate is informing the patient about her elderly mother's stroke:

line		(Candidate 43-1)
12	N:	you know it's a death of a. brain. death of the brain
	...	
17	P:	Well, are you telling me this is irreversible?
18	N:	Yeah. at this ti- . usually at this age .. at 75 years of age
19		it is irreversible . but don't worry . we'll do our best.
	...	
32	N:	Yes, you can talk to her . and er . you know . just tell her
33		that you're here.

34	P:	But you said she's suffered death of the { brain.
35	N:	{Yeah. Death
36		of the brain
37	P:	How can she hear and process information?
38	N:	You know, she can't process any information but .. you
39		the s-, the hearing is the last sense that a patient loses ..
40		just before he dies.

This candidate, to whom a test of English may appear as a simple formality, may have decided (quite justifiably, given the stated purpose of the test, which is made explicit to candidates at the beginning of the oral interaction) that it was more important to demonstrate her complete command of the language than to follow closely the instructions in the roleplay and produce a realistic simulation of how a nurse would behave in real life. Certainly the professional experts were unimpressed by her behaviour - her strategy may not work if they were the judges. The effect on a language-teacher assessor (who in this context may or may not be the interlocutor) of such socially inappropriate interactional behaviour requires further investigation.

POST-WORKSHOP ANALYSIS

While, as previously mentioned, the occupational experts were aware of variation in behaviour, they nevertheless felt that this variation represented the range of possible patient/relative types, and was therefore evidence that the test was valid. From the language tester's point of view, however this view constituted a statement of face validity alone, and might in fact contribute to unfair variability in the challenge presented to candidates. In other words, face validity may mask more serious threats to the content and construct validity of the test. Since the comments made by the nurses in relation to this variability were general in nature rather than focusing on the specific linguistic nature of the variability, the researchers felt it necessary to conduct further analysis of the behaviour of the test interlocutors in order to consider the implications of variability for test takers, interlocutors themselves and the assessment process. This analysis took the form of examination of the recorded workshop interaction, in conjunction with further review of the audio recordings of the test interactions.

The feature of test design that emerged from this analysis as having the most influence on interlocutor behaviour in the OET appeared to be the interlocutor's individual response to the point of tension prescribed in the roleplays. As described earlier, there is an instruction to interlocutors to adopt a particular affective stance, such as anxiety, anger, persistence, nervousness, dissatisfaction, resistance, being upset, or refusal to co-operate. Despite the explicitness of the instructions, and the fact that interlocutors had undergone training, considerable variation was perceived in the extent to which interlocutors did actually adopt the stance prescribed. This in turn was perceived as likely to affect the challenge posed to the candidate, in that, for example, having to deal with a tense, emotional client is considered to be harder than simply imparting factual information.

Specific features of interlocutor behaviour identified in the interactions reviewed in this study were classified into two groups: those appearing to make the test more difficult (which as a rule resulted from the interlocutor throwing him/herself into the role) and those appearing to make the test easier (which generally resulted from the interlocutor behaving in a more teacher-like way). It should also be emphasised that these are perceived as *potential* threats to the fairness of the test and require verification through an analysis of the effect they may have on the scores awarded to candidates, a study which is beyond the scope of this paper.

Features of interlocutor behaviour which appear to make the test harder

1. Sarcasm

One feature of some interlocutors' behaviour, which could make the interaction harder for those candidates who encounter it, is the use of sardonic humour or sarcasm. This was seen as likely to cause particular problems for nurses of a background where such behaviour is not typical, especially in professional or testing contexts. Such behaviour is particularly a concern given that some candidates take the test in their home country while some may have been living in Australia for some time. While sarcasm is a legitimate and realistic reaction to the instructions, the added cultural load involved appears to create an

inequitable challenge for those candidates who encounter it, making them seem less competent in dealing with native speaker patients.

One of the nurses' roleplays (that involving the uncooperative car crash victim) seemed to elicit this type of interlocutor behaviour more frequently than the others. In one test interaction using this roleplay the candidate (nurse) had discovered from questions about the events leading up to the accident that the interlocutor had spent some time in a bar. In response to a question about whether she had drunk any alcohol there, the interlocutor (patient) said, in a sarcastic tone:

'Why do you think I was in the bar?'

Several exchanges later, when asked why she had been drinking, the interlocutor again responded sarcastically:

'Well, I was thirsty, wasn't I?'

The lengthy pause following the first of these sarcastic comments suggested that the candidate was unsure how to respond to the first sarcastic comment produced by the interlocutor. She took it literally; this led to a breakdown in the communication lasting several turns, which the interlocutor eventually had to repair:

'That's why I was in the bar I was drinking of course.'

The second sarcastic comment was also interpreted literally, with the nurse suggesting that the interlocutor could have drunk water in order to quench her thirst.

The point here is not that sarcasm does not represent real-life behaviour in Australia; it most certainly does. However, the problem is one of inequitable distribution: if this is not addressed in interlocutor training, some candidates will encounter interlocutors who use a high level of sarcasm in response to the role-play instructions, while others will not. While it is not certain that a candidate who demonstrates poor ability to respond to sarcasm will necessarily be awarded a lower rating than one of equal ability who is not exposed to it, the potential remains that they *may* be judged as being of different levels of ability. This and the other features of support (and lack of support) require further investigation in relation to scores awarded.

2. Interruption

In interactions where there is a point of conflict (as is the case in the role plays described here) a common strategy is to interrupt, to dispute what the other person says, or not to allow explanations to be completed, in contravention of normal conversational turn-taking rules (see, for example, Levinson 1983). In one of the roleplays reviewed here, the instruction 'be persistent' appears to encourage this sort of behaviour. While such interactional strategies may or may not make the candidate appear less proficient (depending on whether they are allowed to develop their argument fully), they do in any event seem likely to affect the confidence of some candidates, at least, in what is already a highly stressful situation. It has been suggested 'that anxiety inhibits performance for anxious individuals experiencing threatening test conditions' (Covington and Omelich, 1987 cited in Wolf and Smith, 1995:230).

3. Repetition

Another strategy adopted by some interlocutors is repetition either of their view (such as unwillingness to take drugs) or of particular lines of questioning. This appeared to occur more frequently with those roleplays which provide limited stimulus or ideas yet where the interlocutor is expected to extend the interaction for a certain length of time. In such cases, those candidates unable (or unwilling) to demonstrate alternative ways of expressing the same ideas in response are likely to appear as less competent.

One interlocutor adopted a strategy of unwillingness to accept the reason given for her mother's weight loss, as follows:

line		(Candidate 56-1)
78	P:	did she say that to you
79	N:	I told her that . er .
80	P:	mhm
81	N:	er . I will just tell the doctor .
82	P:	mhm
83	N:	er . so I told the doctor .. er . the doctor told me that . er
34		. er he will organise the dentist to see her
85	P:	mhm
86	N:	maybe this afternoon

87	P:	mhm . and do you think she's agreeable ...
88	N:	I haven't talked to her yet (laughs)
89	P:	mhm
90	N:	anyway I'll just talk to her
91	P:	mhm . sometimes it's easier if it come- y'know . sometimes
92		she doesn't want to hear from me .. she doesn't agree with
93		what I say . you know . like mothers and daughters
94	N:	mm
95	P:	sometimes it's better if it comes from somebody . like
96		you . who can . who can .. say that . medically it's better
97		. if she has . er . dentures
98	N:	yes . I will just talk to her
99	P:	mhm
100	N:	don't worry . I will
101	P:	But do you think that's the only reason that she's losing
102		weight? You don't think that there could be some .
103		disease or something eating away in side of her?
104	N:	I just don't know, but I think .. err-. because as you said
105		before . er the food that they served in their old people's
106		house . is .. err . not appea- appealing . maybe that's why
107		she doesn't eat
108	P:	mm . mm . could be . I mean but it's still quite nutritious.
109		I just worry that maybe there's something more wrong
110		with her, You don't think she should go to a specialist or
111		something like that?
112	N:	I think . err .. I should ask . the . opinion of the dietitian
113	P:	mhm
114	N:	err . so that . she will talk to her
115	P:	mhm
116	N:	and . er . ask her what she want
117	P:	mhm . mhm .. So y- .. um-I'm just asking because I'm
118		worried because I'm worried a friend of mine had a . a
119		. relative in a nursing home and the relative lost a lot of
120		weight . and then they found out afterwards that he had
121		cancer . and it was too late to do anything
122	N:	mhm
123	P:	.. an- I . I just wanted to make sure that that's . you don't
124		think it could be that, do you?
125	N:	no-o ... don't worry.
126	P:	mm
127	N:	Er .. do you have a . does she have any problem before?
128	P:	N-nno . no problems just this . you know you can see .
129		she's just skin and bone . you know . she's lost so-o much
130		weight and I'm just . very worried about her
131		(1.0)

132	N:	mm
133		(3.0)
134	N:	she didn't er (1.0) was she (never . have been) in the
135		hospital before
136	P:	no . no . no this is the first time . yeah . yeah. do you think
137		the pneumonia is brought on by the fact that she's so
138		skinny

In this extract, we see a decline in the fluency, length, confidence and accuracy of the nurse's utterances, as the interaction continues without the interlocutor accepting the explanation (ll. 104-107) and proposals (ll. 79-86; 90; 98; 112-116) the nurse offers.

In real life, a nurse certainly might expect to encounter a persistent relative such as this. However, in the testing situation, when the nurse has neither real authority to draw on (since the interaction is ultimately controlled by the interlocutor), nor a reality to refer to (since the patient exists only in the (separate) imaginations of the participants), the nurse's resources may be stretched by this kind of persistence.

4. Lack of co-operation

Some interlocutors adopted an uncooperative role, by failing to respond in expected ways to candidate utterances, and in fact blocking interaction. How candidates cope with this behaviour may depend on their test wiseness, confidence, creativity or on how much professional experience they are able to draw on, as much as on their language proficiency. For example:

line	(Candidate 56-2)	
22	N:	Good morning, I'm the nurse assigned to take care of you
23		today . hh . and ... um . according to our report this
24		morning . that you refused to take .. the pain . pain
25		reliever .. er the um medicine .. to relieve the pain
26	P:	I just want people to leave me alone
27		(1.4)
28	N:	Why, what happened ..
29	P:	Well , I just don't want people asking me questions all the
30		time I just want to be left alone
31		(1.0)
32	N:	Don't worry, I can talk to you as a friend, not as a nurse
33		.. er you can tell me anything

34	P:	I can't I can't. I just don't want to, I've got a hangover
35		.. I - I feel awful, my head aches .. my arm aches . and I
36		really just don't want to talk to anybody
37		(2.4)
38	N:	I can, I can help, if you'll tell me what . what you're . what
39		you're feeling
40	P:	I'm feeling like I've got a terrible pain in the head . and
41		I'm feeling really unhappy . {and angry
42	N:	{mm

In this instance the candidate appeared to cope rather well with this problem, in that she found four different approaches to persuading the interlocutor to agree to talk, the last of which proved successful.

Features of interlocutor behaviour which appear to make the test easier

1. Factual questions

Sometimes interlocutors introduced a variety of factual questions into the interaction (for example, requesting explanations of simple medical terms or procedures), rather than pursuing the point of tension prescribed, thus allowing the candidate to use straightforward language functions such as description and advice, and in some cases allowing her to use the interaction as an opportunity to display professional knowledge, for example:

> (Candidate 43-1) 'but nurse, what would have caused this?' (i.e. the stroke)

A longer extract demonstrates how one interlocutor conducts the interaction largely as a questioning routine without really attending to or following up on the responses. She does not become 'increasingly agitated' as the roleplay card instructs, rather the reverse, if anything.

line	(Candidate 66-1)	
22	P:	What's, {what's
23	N:	{she's now, she's now unconscious but errr, but
24		er it err er it it er it stopped it will stop er . very soon
25	P:	You think that she'll regain consciousness soon?
26		(0.8)
27	N:	err she'll she will conscious yes I {think so
28	P:	{ a-and

what what do you think is likely to be the result? I'm really worried. Is she going to be paralysed? Is she going to be able to speak?

N: I hope, I hope err she will be all right. She will be healthy again and err there are er you know er any erm effects of this (drug)

P: Is there any way of telling while she's still unconscious?

(0.7)

N: Umm (0.5) No, no .. no . she will be OK don't worry about it and err... um ..

P: Can you tell from her breathing or from the blood pressure if the situation is bad or not bad I mean

(0.5)

N: er

(0.5)

P: it seems .. it seems so difficult to tell when she just lying there like that

(0.5)

N: Yes. It's difficult to see her like she is but . here, here .. err . she is . she has normal pulse she has normal blood pressure, it should be OK and, and, and, er, don't worry about it and, er . it takes sometimes but finally it will {be (???)

P: {Is there any way you can resuscitate her more quickly, that she can be brought back to consciousness?

(2.0)

P: Or is it just a question of waiting?

(0.8)

2. Linguistic simplification

This kind of accommodation is common amongst interlocutors (Ross and Berwick 1992; Ross 1992; Morton, Wigglesworth and Williams, forthcoming). Ross and Berwick's findings indicate that in general simplification (or *accommodation* in their terms) is based on interlocutors' professional judgement of candidates' language proficiency, being more frequent with weaker candidates than with stronger ones, although they do find instances of under- and over-accommodation.

Types of interlocutor simplifications found in this study included repetition, reformulation/syntactic simplification, slowing of speech, and over-enunciation, sometimes in combination, for example:

(To candidate 40-1)

> ''I was mowing the lawn - cutting the grass'

(To candidate 66-1)

> 'Is there any way you can resuscitate her more quickly, that she can be brought back to consciousness?'

(To candidate 66-1)

P: What's going to happen now? What's going to be done? You say you're checking her pulse, {her blood
N: {yes yes
P: pressure. What happens next?

The use of simplification by interlocutors was remarked upon more than once by the workshop participants although they were generally unable to specify the particular linguistic behaviours:

> - 'I think the interlocutor, she sort of led her quite a lot, actually I found that where the nurse was supposed to reassure here, that she, the other one was always taking control: it was things like, 'Oh well, look, I'll speak to the doctor - that sort of thing, so I mean it wasn't really reassuring'
>
> - 'She (i.e. the patient) was in control, I felt'
>
> - 'And the aggression, she was not really aggressive or demanding, it all abated, because you know the, the nurse was struggling along there, she was being too kind'.

The longer extract on the preceding pages (Candidate 66-1) is taken from the interaction which elicited this last comment. Not only does the interlocutor reject the confrontational role prescribed in favour of question-and-answer, but she appears to modify her questions, presumably as a result of her developing impression of the candidate's lack of ability. These modifications include repetition and reformulation, elaboration, lexical simplification.

While it is likely that this kind of behaviour is frequently motivated by the perceived lack of proficiency of the candidate, it is unclear to what extent this is always the case and to what extent it is a trait of certain interlocutors. The workshop participants were able to identify certain interlocutors as consistently simplifying their language with different candidates, regardless of their ability.

3. Allowing the candidate to control the interaction

In some instances, the interlocutor lets the candidate initiate topics and control the direction of the interaction, rather than following the instructions to be persistent or assertive. This can also make for an easier challenge, allowing the candidate to show off her knowledge, for example:

line (Candidate 43-1)

4 N: All right..er-oh. I'd like to tell you something about your mum..
5 P: Yeah..
6 N: You know your mum's already 75 years of age..
7 P: yes..
8 N: an er. she's had what we call a. a. cardiovascular accident.
9 P: yes
10 N: Something to do with the brain.
11 P: Right

It appeared on the basis of the sample of interactions reviewed here that some interlocutors seemed to adopt this behaviour more consistently than others. This suggests that interlocutors position themselves on a continuum from identification with the role (more patient-like behaviour) prescribed to identification with the candidate (more teacher-like behaviour). From the limited sample of data examined here, it appears that some interlocutors are unable or unwilling to step outside their professional role, whereas others appear to adopt the prescribed roles without difficulty or discomfort. It is possible that still others may vary their behaviour according to their perception of each candidate's proficiency. It may also be that they vary their behaviour non-systematically, but in response to how comfortable they feel with particular roleplays. These hypotheses require verification through discourse analytic studies of interactions involving a range of interlocutors, candidates and roleplays.

The behaviours described above, both those perceived as contributing to difficulty as well as those appearing to support the candidate, illustrate the wide range of interactional strategies that interlocutors select from. While the task instructions introduce the element of emotional complexity, and the need to negotiate, persuade, etc., interlocutors respond to them in very different ways: in general it

seems that those who follow the instructions most closely add extra difficulty, while those who choose not to behave in this manner provide a more straightforward challenge for the candidate.

CONCLUSIONS

This study has considered some of the issues associated with transferring occupational interaction to a testing situation. The value of this general approach to language testing derives from its focus on contextualised language use, reasonably representative of real-life interaction. Findings suggest that it is possible to simulate occupational communication tasks reasonably effectively, but that this is not a straightforward process. Of the four roleplay materials considered in this study, two were found to present no problems of inauthenticity of content, while the other two were seen to exhibit a variety of problems, despite their having been developed in conjunction with professional experts. The problems arose not so much from the inclusion of incorrect information, but more from what appeared to be the result of the need to be as concise as possible in the information and instructions presented to the participants. In the materials reviewed, this brevity led at times to ambiguities and the exclusion of necessary detail.

The fact that the problems in the test materials really only came to light during a review of test interaction indicates the inadequacy of reviewing written materials only. Specific-purpose test materials such as these require input from relevant occupational experts. However, such people are generally not familiar with the constraints of the testing situation, and are unlikely to be aware of the problems that might arise for candidates as the result of transferring their (the writers') visualisation of a real-life context to a few lines on a role play card. Similarly, the test developer is unlikely to be familiar with the occupational context and may not be able to identify problems in relation to content. We propose that one step in the test development process should involve (as it did here for test validation purposes) the two parties reviewing jointly the appropriateness of written role play materials *in conjunction with* recorded samples of test performance. Only in this way can the effectiveness of the materials be gauged, and specific problems be identified.

Why is it so important to ensure that the materials lead to 'authentic' interaction in a test whose purpose is not to assess occupational competence? The answer to this is simple: fairness to candidates. Where candidates are required to draw on skills other than language (in this case background knowledge and strategic competence) in order to perform on the test, it is vital to ensure that all candidates are presented with a similar challenge. If some roleplays demand more creativity or fast thinking to deal with the ambiguities or missing information, then this is going to create unequal challenges across tasks. If some candidates are more 'test wise' than others, more aware of what exactly is being assessed, or more confident, they are likely to perform more convincingly. While the actual effect of the performance on the assessor is unknown, it is certainly likely to be the case that a confident candidate is will present a 'good' sample of performance, whereas one who is confused or uncertain about what to say is less likely to.

Test writers, furthermore, need to be aware of the effect on the emotional tone of the interaction of instructions to interlocutors. As we observed here, these instructions are not all of the same order; some lead to potentially much more challenging communicative demands, again raising the issue of fairness. Whether the materials writers are occupational experts or language teachers or testers, again, we propose that the effect of the instructions be monitored through a review of actual performance as part of the test development process.

In addition, it is clear that interlocutors require substantial *training* in order for them to adopt the role envisaged for them by the test developers and the materials writers. They need to be made aware of the wide range of possible interlocutor behaviours and the effect that different strategies may have on outcomes. They then need to be informed as to how they should interpret the instructions. Variation cannot, of course, be eliminated. To do this would reduce a performance test to a series of predictable exchanges: it would cease to be a test of ability to use language in interaction. However, the variability can and should be substantially reduced where it is likely to affect candidates' chances in tests which carry significant consequences for candidates.

While interlocutors for the OET are required to take part in training sessions, it is perhaps the case that these sessions focus more on reliability of ratings and not enough on consistency of interactional techniques. In

our experience this is typical of language test training, and reflects a concern for reliability of ratings at the expense of reliability of elicitation of that performance. Both rater behaviour and reliability are relevant to the outcomes for the candidate, and both should receive adequate attention; initial training should ideally be combined with on-going monitoring and feedback. Tests typically require the interaction to be recorded for monitoring standards of assessment; there is no reason why these recordings should not be used additionally to monitor interlocutor behaviour (see Morton et al (forthcoming) for a description of how this is achieved in relation to a test for intending migrants to Australia). The point here is that it is not possible to set up 'real-life' situations and assume they will all be 'real-life' in the same way (and where interlocutors behave in a particularly 'teacherly' manner, they may be quite unlike 'real life'). Our aim here is not to bring into question the use of performance tests based on tasks which simulate real-life professional interaction (to do so would imply a return to indirect methods of testing with all their attendant questions of validity), but rather to identify ways of improving practice.

Ultimately the findings of studies such as this (as well as those mentioned earlier - Caffarella, 1994; Filipi, 1994; and Lazaraton and Saville, 1994) are of concern only insofar as variations in interaction have an impact on the judgement made of candidates' performance. If it can be shown that particular styles of interaction systematically lead to either lower or higher scores for test candidates, then the validity of the assessment procedure is called into question. Recent studies are beginning to indicate that what goes on between the two participants in an oral interview is in a complex relationship with the score awarded. A study by McNamara and Lumley (1993) found that scores were affected by the perceived 'rapport' between interlocutor and candidate. Surprisingly, a less than excellent rapport between the interlocutor and candidate led to a small advantage to the candidate in the scores given. In the study by Morton et al. (forthcoming) a similar finding was reported in relation to the perceived 'competence' of interlocutors. The present study does not seek to address the relationship between performance conditions and test scores; rather it should be seen as a preliminary study which aims to identify factors which have the *potential* to affect test scores. Follow-up studies using rated and analysed samples of test interaction will be required to investigate whether these factors do affect scores, to determine

which of them pose *real* threats to validity and which are examples of acceptable variation.

NOTES

1 This study was conducted with the assistance of a Special Initiatives Grant from the Faculty of Arts, The University of Melbourne. We wish to acknowledge the contribution of the National Languages and Literacy Institute of Australia (NLLIA) in supporting the ongoing research associated with the Occupational English Test, and of Elisabeth Grove for her assistance with materials used in this study.

2 These roleplay cards are included as an example of the materials used in the test. It is not possible to include other roleplay cards for reasons of test security.

3 A CVA is a cerebro-vascular accident, more commonly known in lay terms as a stroke. It is a standard abbreviation in health care.

4 Each audio recording contained the interaction of one candidate; each candidate completed two tasks randomly selected from the same set of four tasks.

5 For reasons of confidentiality, it is not possible to reproduce transcripts of all interactions referred to in this paper.

6 The codinbg system used throughout the paper indicates the candidate number (e.g. 69) and whether the role play is the first or the second attempted by the candidate. Sometimes the utterances of only one particular participant are included.

7 The coding system used throughout indicates the candidate number (eg 75) and whether the roleplay is the first or the second attempted by that candidate; sometimes the utterances of only one participant are included.

REFERENCES

Bachman, L.F. (1990) *Fundamental Considerations in Language Testing*. Oxford, Oxford University Press.

Brown, A. (1994) LSP testing: the role of linguistic and real-world criteria. In Khoo, R. (ed.) *LSP: Problems and Prospects*. Singapore, SEAMEO Regional Language Centre.

Brown, A. (1995) The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, 12, 1: 1-15.

Cafarella, C. (1994) Assessor accommodation in the 1992 Victorian Certificate of Education Italian oral test. Unpublished MA thesis, University of Melbourne.

Cumming, A. and R. Berwick (1996)*Validation in language testing*. Clevedon, Avon, Multilingual Matters

Davies, A.D. (1986) Indirect ESP testing: old innovations. In Portal, M. (ed.) *Innovations in Language Testing*. Windsor, Berks., NFER-Nelson.

Davies, A. (1995) Testing communicative language or testing language communicatively: what? how? *Melbourne Papers In Language Testing* 4,1:1-20.

Douglas, D. (forthcoming) *Field-specific language testing*. Unpublished manuscript.

Douglas, D. and L. Selinker (1993) Performance on a general versus a field-specific test of speaking proficiency by international teaching assistants. In Chapelle, C. and D. Douglas (eds) *A new decade of language testing research*. Alexandria, VA, TESOL Publications.

Elder, C. (1993) How do subject specialist construe classroom language proficiency? *Language Testing* 10, 3: 235-254.

Elder, C. (1994) Are raters' judgements of teacher proficiency wholly language based? *Melbourne Papers In Language Testing* 2,2:41-59.

Filipi, A. (1994) Interaction or interrogation? A study of talk occurring in a sample of the 1992 Victorian Certificate of Education Italian oral common assessment task. Unpublished MA thesis, University of Melbourne.

Lazaraton, A. and N. Saville (1994) Processes and outcomes in oral assessment. Paper presented at the Language Testing Research Colloquium, Washington, April.

Levinson, S. (1983) *Pragmatics*. Cambridge, Cambridge University Press.

McNamara, T.F. (1990) Assessing the Second Language Proficiency of Health Professionals. Unpublished PhD Thesis, University of Melbourne.

McNamara, T.F. (1996) *Measuring Second Language Performance*. London, Longman.

McNamara, T.F. and T. Lumley (1993) The effect of interlocutor and assessment mode variables in offshore assessments of speaking skills in occupational settings. Paper presented at the Fifteenth Language Testing Research Colloquium, University of Cambridge, August.

Morton, J., G. Wigglesworth and D. Williams (forthcoming) Approaches to the evaluation of interviewer behaviour in oral tests. In Wigglesworth, G. and G. Brindley (eds) *access:: issues in English language test design and delivery*. Sydney, NCELTR.

O'Loughlin, K. (1995) Test taker performance in direct and semi-direct versions of an oral proficiency test. Unpublished manuscript.

Ross, S. (1992) Accommodative questions in oral proficiency interviews. *Language Testing* 9,2: 173-186.

Ross and Berwick (1992) The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition* 14:159-176.

Wesche, M.B. (1987) Second language performance testing: The Ontario Test of ESL as an example. *Language Testing* 4,1: 28-47.

Wesche, M.B. (1992) Performance testing for work-related second language assessment. In Shohamy, E. and R. Walton (eds) *Language assessment for feedback: testing and other strategies*. Washington, D.C., National Foreign Language Center Publications.

Wolf, L.F. and J.K. Smith (1993) The consequence of consequence: Motivation, anxiety and test performance. *Applied Measurement in Education* 8,3:227-242.

Young, R. and M. Milanovic. (1992) Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition* 14: 403-424.

RASCH SCALING: HOW VALID IS IT AS THE BASIS FOR CONTENT-REFERENCED DESCRIPTORS OF TEST PERFORMANCE?

Joy McQueen
Australian Council for Educational Research

ABSTRACT

This paper describes a methodology for exploring the validity of the Rasch scaling procedure in relation to a multiple-choice test of reading Chinese as a foreign language. The validation procedure involved a post hoc content analysis of test items to identify factors which the research literature on early reading indicated were likely to be related to item difficulty. A comparison was then made between these difficulty factors and the item-difficulty ranking produced by a Rasch analysis of test items. The analysis revealed that the dimension mapped out by the Rasch scaling (and to some extent the corresponding wording of the reporting descriptors) does indeed reflect certain elements thought to be related to early reading ability.

INTRODUCTION

Over the last decade there has been a growing interest in the use of the Rasch (single parameter) school of Item Response Theory (IRT) as a tool to study content domains, and to define dimensions of ability (see for example Mossensen et al. 1987; Ballas 1987; Australian Council for Educational Research [ACER], 1989). In this procedure, common scaling of items and ability produces 'maps' of candidate ability/item difficulty, and allows content-referenced interpretation of candidates' performances. Each level of performance 'is defined by describing the types of...question that most students attaining that proficiency would be able to perform successfully' (Mullis and Jenkins 1988: 141). Descriptors make test results meaningful to test candidates and the general public in a way that scores do not, but their validity inevitably rests on the accuracy of our understanding of what is actually being tested by the items. This understanding cannot be taken for granted. As Wood (1991: 16) puts it, it is 'right to say that understanding what makes an item difficult is still something we know too little about.'

ARAL Series S Nº 13 (1996) 137-187

Rasch scaling is used as the basis for reporting student performances on the Australian Language Certificates (ALC) tests[1]. Results of the tests are reported through descriptions of typical skills at each of three levels of performance. The ALC programme offers tests of listening and reading for beginning learners (ie with up to 100 hours of instruction) in French, German, Indonesian, Italian, Modern Greek, Japanese and Modern Standard (Mandarin) Chinese for lower secondary school students. In 1995, 56,000 students in over 1000 schools in Australia, New Zealand and the Pacific participated. Because of the large numbers of participants and the need for rapid and inexpensive scoring, the tests use a multiple-choice, machine-markable format.

This paper focusses on the 1990 ALC test of Chinese reading and describes a methodology for exploring the validity of a Rasch-derived scale and the content-referenced descriptors based on it. The major steps in the validation procedure are as follows:

1 Identification of possible predictors of reading difficulty;
2 Post hoc analysis of test items, relating skills tapped by the items to the predictors identified in step 1;
3 A comparison of these difficulty factors with the item-difficulty ranking produced by the Rasch scaling procedure;
4 Consideration of the possible influence of task-type on item difficulty;
5 Examination of the ability-level descriptors with respect to factors arising in steps 1 to 4.

THE READING PROCESS

In order to examine the way in which the ALC Chinese reading test measures foreign language reading ability, it is necessary to consider the nature of reading itself.

In recent decades there has been a move away from the traditional view of reading as simply a 'bottom-up' or decoding process. Goodman's view of reading as a 'psycholinguistic guessing game' (1976) has been widely accepted. According to his 'top-down' model of reading, higher-level conceptual processes are most important, especially for good readers. This model represents reading as a process in which the reader

makes sense of the text by using contextual clues to predict what will come next, and sampling the text to test the resulting hypotheses. In this model, syntactic and discourse-level features are of particular importance. The findings of a study of upper secondary learners of Chinese as a foreign language supported this view, showing that, for the intermediate-level group studied, 'knowledge of discourse and syntactic structure is more significant in reading comprehension than is recognition of characters' (Pavlidis 1990: 17).

According to some theorists (LaBerge and Samuels 1974; Schneider and Shiffrin 1977; Shiffrin and Schneider 1977), higher-level processing requires that lower-level processes be fully automated. When word recognition is not automatic, conscious attention to the decoding process diminishes the capacity available for higher cognitive processes.

Interactive views of reading (Rumelhart 1977) have gained currency, perhaps partly because empirical research has not always supported the top-down model (Stanovich 1980). Reading is seen as a process in which higher-level and lower-level cognitive processes occur simultaneously and inform one another. One form of the interactive approach is schema theory (Rumelhart 1980), according to which experience and knowledge of the world provide a framework for interpreting what is read. Our expectations of the content and form of a text, and hence our understanding of it, are shaped by our previous experience.

Stanovich (1980) extended the interactive model, describing an 'interactive compensatory' process in which information at one level of processing would compensate for absence of information at another. Thus, readers with poor word-knowledge were seen to rely heavily on contextual clues for meaning. It must be said that the interactive-compensatory model, like almost every other model, has its critics (e.g. Freebody and Anderson 1983). Moreover, attention is again being paid to the role of decoding skills (see for example Pino-Silva 1993). There is also evidence that, at low levels of second language proficiency, reading is basically a bottom-up process, with readers attending primarily to graphic information (Cziko 1980; Hatch 1974)

The test considered here is designed for elementary-level students, and it is therefore appropriate for the remainder of this paper to concentrate on lower-level skills of reading. This is not to deny the place

of higher-level skills in the reading process, but in beginner-level processing of a foreign language text, decoding necessarily plays a primary role.

Reading comprehension and word knowledge

> To read, a reader must know words. To become a better reader, a reader must learn more words...improvement in reading comprehension can be directly attributed to an increase in vocabulary
>
> (Hague 1987: 218).

At least since Davis's 1944 study, word knowledge has been seen as closely related to reading comprehension. This relationship might be interpreted in a number of ways. Anderson and Freebody (1981) present three views:

(a) the instrumentalist position, which interprets the link as a causal one, i.e. that it is word knowledge which enables reading comprehension;
(b) the aptitude position, according to which reading comprehension and word knowledge share a single underlying ability; and
(c) the knowledge hypothesis, which, like schema theory, suggests that word meaning is a function of its context, and understanding of the word implies familiarity with the current topic.

Anderson and Freebody conclude that 'it would be naive to assume that one of these positions will turn out to be entirely right and the other two entirely wrong' (1981: 82) — a reasonable conclusion since, at least in the case of the instrumentalist position, empirical evidence is contradictory. The review by Anderson and Freebody reports research results which are mixed, but which on the whole do not support the instrumentalist position. The more recent set of findings discussed by McKeown and Curtis (1987) is even more negative. However, whatever the nature of the role played in reading by word recognition, there can be no doubt that it is vitally important. 'Mastery of rapid, context-free word identification is one of the major factors that separates good from poor readers' (Juel 1983: 307).

In addition to its importance in ordinary reading, in a testing context word recognition can take on extra significance. The processing of small chunks of text might be encouraged by the nature of the items. Some test tasks might not be authentic, and might require that candidates perform activities, such as word-matching and translation, which are not part of the normal reading process. To the extent that this is true of a particular test, the influence of word knowledge on the test results will be artificially inflated.

Readability and reading comprehension

'The concept of readability in a general sense is implicit in all discussions of comprehension' (Williams and Moran 1989: 218). Top-down and interactive approaches to reading have proposed that meaning does not reside solely in the text, but is constructed by the reader in terms of his or her experience. Since no reader can construct meaning from a text which is linguistically inaccessible, it is pertinent to consider the factors that might make a text inaccessible to less competent readers. Indicators of readability may be possible predictors of reading comprehension, and therefore of item difficulty.

For several decades readability formulae enjoyed considerable popularity. According to Alderson and Urquhart (1984: xxii) the most reliable predictors of readability were found to be word difficulty (often defined in terms of frequency and length), and sentence complexity (often indicated by sentence length). Chall (1987: 11), in summing up readability research of the previous sixty years, offers as the most reliable predictor 'vocabulary difficulty (as measured by word familiarity, word frequency, word length in syllables or letters, abstractness of words or difficulty of concepts)'.

The issue of familiarity is even more relevant for character texts than for texts written in a romanized form such as Pinyin[2]. As an alphabetic script, Pinyin provides orthographic clues to the meaning of unknown words. Characters, although ideographic in origin, no longer offer reliable pictorial clues to their meaning. Some have phonetic components which can be of some use in deriving meaning, but a working knowledge of these is usually built up only after considerable exposure to the script.

The word-frequency factor is more relevant to readability in a first language than in a foreign language which is being learned as a school subject where frequency of use is determined artificially by constraints imposed by the syllabus. The degree of correspondence between classroom frequency of use and real-life frequency is uncertain. As an indicator of exposure to particular words included in test items it is more appropriate to consider whether the words feature in the syllabus, and in what order they are presented.

Laroche (1979) recommends that readability formulae for foreign language texts should pay greater attention to syntactic complexity than do readability formulae for first language texts, since syntactic complexity is more likely to be an obstacle for a reader negotiating a second language. He suggests that sentence length might be used 'as a crude attempt at quantifying syntactic complexity' (1979: 135).

For the test in question, however, sentence length cannot be used as an indicator of syntactic complexity because (unsurprisingly, since this is an elementary-level test) two of the six stimulus texts are written in short phrases and sentence fragments. Instead, some indication of syntax-related difficulty will be obtained from counts of syntactic patterns which are unlike those of English.

Yang (1971), in developing a readability formula for Chinese as a first language, found that of the 39 independent variables investigated, the most reliable predictors were word frequency, character stroke-count and sentence length. At first glance the inclusion of the average number of strokes per character appears to be at odds with the finding by Hayes (1987) that character complexity had no significant effect on the recognition of characters by native or non-native speakers. The apparent contradiction might be explained in terms of the nature of the relationship of readability and reading comprehension. Readability indices are intended to *predict* reading comprehension. There is no suggestion that the factors which they highlight actually *cause* variations in the difficulty of reading texts. Although the precise nature of the relationship between character complexity and text difficulty is unclear, on the basis of Yang's finding, character complexity will be included in our array of potential predictors of reading difficulty.

ASPECTS OF SECOND OR FOREIGN LANGUAGE WORD DIFFICULTY

The state of research

To this point, a number of possible predictors of difficulty in reading comprehension have been enumerated. We now turn to focus more closely on second/foreign language word recognition, and factors which might facilitate or impede it. Despite a general acknowledgement that semantic processing is central to both first and second language reading, little is known about the differences between first and second language word knowledge (Koda, 1994: 10). A number of possible sources of second or foreign language vocabulary difficulty have been identified by research in the field although empirical and theoretical findings are surprisingly sparse (Meara 1983).

Of the studies that do exist in this area a large proportion have been psycholinguistic in orientation, dealing with lexical storage in bilinguals. A second major category of interest has reflected pedagogic concerns. Important considerations here have been the criteria for construction of word-lists (see for example Richards 1974), the nature of 'coreness' (Carter 1987; Dollerup, Glahn and Hansen 1989), and the effectiveness of methodologies for teaching the word-lists (for example Raugh and Atkinson 1975). There has been little overlap between the psycholinguistic and applied linguistic traditions in the field of vocabulary research (Meara, 1987: 2; 1993: 762) and there is a paucity of empirical research bridging the two. Even less deals directly with the issues which concern this study with studies ranging from strategies for inference of unknown words (Palmberg 1987, 1988) to the measurement of vocabulary knowledge per se (Read 1988). Studies of the relationship of reading and second/foreign language vocabulary acquisition are most often concerned with advanced-level students (see for example Dubin 1989; Huckin, Haynes & Coady 1993).

Intralexical factors in second language vocabulary difficulty

In 1990 Laufer reviewed studies of second language vocabulary acquisition and compiled a list of potential sources of difficulty. In this section, the relevance of the components of her framework for this study

are considered. Those which are irrelevant are discarded, and other relevant factors identified elsewhere are incorporated in order to produce an inventory of possible predictors of reading difficulty. Following this, a content analysis of the ALC Chinese test in terms of the newly developed framework will address the issue of the validity of the mapping procedure and the descriptors which are based on this procedure.

Laufer categorizes possible intralexical difficulty factors as phonological, grammatical, semantic, register and 'multiple meaning' factors (see table 1). Each of these is discussed below.

Phonological factors	
	length
	pronounceability
Grammatical characteristics	
	part of speech
	inflectional complexity
	derivational complexity
Semantic features	
	abstractness
	specificity
	idiomaticity
Register restrictions	
	neutrality
Multiple meaning	
	polysemy and homonymy

Table 1 *Possible intralexical factors* in word difficulty (abstracted *from Laufer 1990)*

Phonological/Word-form factors

Although Laufer labels the factors in this category as 'Phonological' 'Word-form' is perhaps a more appropriate label as reading is the skill

under discussion. The two factors, length and pronounceability, are considered in turn.

In Chinese, whether written in characters or in the Pinyin romanized script, the measurement of word-length is complicated by uncertainty about word boundaries. A character represents a single morpheme. A word may be composed of one or more morphemes, but this fact is not reflected in differential spacing of characters. Word boundaries are not visually marked in character texts. In Pinyin script, where word-boundaries are formally marked by spacing, there are few conventions for deciding what constitutes a semantic unit. Is the expression for *Monday* one word *(xingqiyi)* or two (*xingqi yi*)? The Australian Language Certificate examination paper did not always observe the same spacing as the major textbook used by Chinese students at this level, *Hanyu* (Chang, Mackerras and Yu 1990). In this study, the boundaries of words written in Pinyin were taken to be those applying in the test paper. Boundaries for character text words were those observed by the textbook *Hanyu*, when it presented vocabulary in Pinyin form.

With respect to pronounceability Laufer reports research which generally supports the view that pronounceability (conceived in terms of phonological patterning and its similarity to, or difference from, patterns of the first language) affects both production and reception of a new word. While intuitively this factor might appear relevant to listening rather than reading, an extensive review (Gibson and Levin, 1975: 201-206) of first language reading studies suggests that pronounceability, or 'legal orthography' may be influential in reading performance.

In Pinyin texts the initial consonants *q* (with no following *u) zh*, and *x* represent Mandarin phonemes which are quite unlike any English pronunciations of *z*, *q* and *x*. Encountering these instances of 'illegal' initials can be daunting for beginning students, and whether or not the perceived difficulty of pronunciation exceeds the actual difficulty, there is potential for some impediment to the recognition of words. The extra difficulty generated by tonal languages (for speakers of non-tonal first languages) should be mentioned here. The difficulty is perhaps more general than word-specific. That is to say, pronunciation is generally made more difficult by the need to observe tone values, but this will not account for variations in difficulty between one Chinese word and another.

Grammatical characteristics of the word

Three factors contribute to grammatical characteristics: part of speech, inflectional complexity and derivational complexity. While Laufer reviews findings which claim that some parts of speech are easier to learn than others these will not be considered here as methodological problems render them inconclusive (Laufer, 1990: 298). Likewise, because there is little inflectional morphology in the Chinese language, this issue will not be pursued.

Derivational complexity of individual words, on the other hand, may well have an effect on the difficulty of those words. According to Laufer (1989), analysis of a word into known morphemes may assist recognition, but there is always the danger of 'deceptive transparency': analysis of a word into its components can be misleading in some cases. Nation (1990:36) also highlights the predictability of meaning as a possible influence on the 'learning burden'. He raises the issue of patterns of use of second language words and suggests that the predictability of patterns of use, and their similarity to first language patterns, may influence the ease with which words can be learnt. This may be particularly important for a language like Chinese, which is relatively distant from the learners' first language and thus the presence of syntactic patterns unlike those encountered in English will be included as a potential predictor of difficulty.

Semantic features

There are three semantic factors: abstractness, specificity and idiomaticity. Ludwig (1984: 557) holds that abstractness hinders vocabulary learning. However, Laufer (1990: 300) claims that concreteness is not in itself a facilitating factor. Because characterisation of a word as 'concrete' or 'abstract' can be problematic, (see Allen & Valette cited in Laufer 1990) the influence of abstractness is not investigated here.

Similarly specificity is not examined. While Blum and Levenston (1978: 152) argue that learners favour superordinates while native speakers make more use of co-hyponyms, from which Laufer extrapolates that 'foreign learners retain the general items better than the specific ones' (1990: 301), the application of the general/specific distinction is difficult at beginner levels. This is because the pool of vocabulary items on which

the test can draw is too small to allow effective use of this kind of categorisation. Likewise, while Laufer reports several studies which confirm that idiomaticity adds to word difficulty in a second language, the use of idiom, being largely precluded by the elementary level of the test, will not be considered further in this paper.

Two factors not mentioned in Laufer's taxonomy are codability and affect. Affect is excluded from further discussion since the range of vocabulary covered by the test is so limited. However, codability is addressed. Higa suggests that 'foreign words whose referents are highly codable in the learner's language would be learned more easily than those whose referents are low in codability' (1965: 171). Nation (1990: 36) makes a similar point. As MacWhinney (1995: 306) says, 'in the very worst case, learning of a new word would involve a...totally unfamiliar and complex set of semantic meanings'. Certainly there is intuitive appeal in the idea that where learning a word involves forming an entirely new concept, the process will be more difficult than for words which have approximate equivalents in the native language.

Register restrictions

Halliday, McIntosh and Strevens (1964: 88) argue that register confusion is an important source of error for second language learners. Laufer claims that 'neutral words, which can be used in all registers, will be easier to learn' (1990: 302). However, this claim does not lend itself to investigation because of the restricted range of vocabulary covered by the test. Thus the effect of register restrictions will not be considered here.

Multiple meaning

Laufer (1990: 303) offers evidence of the influence of polysemy and homonymy on both receptive and productive vocabulary. Henning's (1973) study of the second language lexicon provides evidence that, for beginners at least, new words are encoded in the memory according to sound, and similarity of sound is therefore a source of confusion. This is compounded by the large number of homonyms in Chinese. As mentioned earlier, word boundaries within character texts can be unclear to the low-level reader. This fact, combined with the considerable functional overloading of Chinese morphemes (Pavlidis 1990: 3) can

make it difficult for the beginning reader to determine which of several possible grammatical functions is being performed by a character in a particular environment (Pavlidis 1990: 3). As Pavlidis points out,

> while the ability to recognize individual characters is clearly necessary, a knowledge of the grammatical function of characters in a given context is crucial to assigning appropriate meanings to those characters and to determining their functional and semantic relationship to the other characters in the sentence.

The problem also exists to some extent to Pinyin script. As previously noted, Pinyin word boundaries are not fixed, and the boundaries observed in the test paper differed in some cases from those used by the text-book. It seems, then, that word recognition in Chinese character texts involves tasks of syntactic and semantic identification that might not occur in the reading of alphabetic scripts.

Script

Nation (1990: 36) includes the script-type as a possible source of difficulty. Pinyin is an alphabetic script which offers phonetic and semantic clues to word meaning. As noted earlier, characters offer little assistance to the beginner in this respect. Furthermore, while character-text spacing is based on the morpheme, Pinyin spacing is based on word-units. Intuitively one might therefore expect that English speakers would have more difficulty with a passage written in Chinese characters than with an equivalent passage written in Roman script. Everson (1988) provides evidence to this effect, showing that the (low-level) learners in his study did rely heavily on the semantic and phonemic clues provided by Pinyin.

Extralexical factors

Laufer's review was concerned with *intralexical* factors in word difficulty. However, there is no doubt that learner factors are also significant. Indeed, Higa goes so far as to say that 'it is chiefly the learner's past and present learning experience that makes any learning material easy or difficult' (1965: 168). A most important element of learner experience must be the degree of exposure to the words in question (Higa 1965: 175; Nation 1990: 43-44). The only available indication of the relative familiarity of vocabulary items is the order of

presentation of vocabulary in *Hanyu*, the textbook used by most of the students. This is discussed in more detail below.

Both Higa (1965: 173) and Nation (1990: 45-46) refer to the increased learning burden which, in their view, results from the common practice of grouping words with synonyms and antonyms for teaching purposes. Tinkham (1993) goes further, providing evidence that the learning burden is increased when words are grouped into lexical sets for presentation to the learner. 'Grouping for presentation' is added to our list of predictors.

Laufer's original table can now be modified by the removal of factors which are not applicable in this study, and by the addition of others drawn from other sources on readability and second language word recognition. The resulting list (table 2) outlines the revised inventory of possible influences on item difficulty in the ALC reading test:

Word-form factors
length
pronounceability
Grammatical characteristics
derivational complexity
syntactic patterns
Semantic features
codability
Multiple meaning
polysemy and homonymy
Written form
script-style: pinyin vs characters
character complexity
Extralexical factors
order of presentation
presence of unknown lexical items
grouping for presentation

Table 2 *A modified list of possible influences on word difficulty in Chinese reading passages*

TEST METHOD EFFECTS ON ITEM DIFFICULTY

In addition to the factors listed in table 2, an important potential source of item difficulty, external both to the candidate and to the language of the test material is the style or type of test item employed. Bachman points out that research into test method effects has not been sufficiently detailed, tending to treat them as 'monolithic wholes' (1990: 115). The list of possible method effects is enormous. Bachman's (1990: 119) framework is comprehensive, and includes categories of test facet which might account for differences in item difficulty. Of these, perhaps the most obvious potential influence on item difficulty is the task type. This facet will be further considered later in this paper.

THE TEST

Five hundred and ninety two secondary school students in Australia and New Zealand participated in the Chinese test. Of these 80% named English, and 5% Chinese, as the language spoken at home. The test was open to students who were in their second year of learning Chinese, or who had received between 80 and 120 hours of instruction. The year-level break-down of participants is given in table 3:

Year	Percentage
8	26
9	60
10	10
11	2
12	1
missing	1

Table 3: Participants: Chinese ALC 1990 (Source: ACER, 1990:5)

The question papers (Appendix 1) consist of 25 multiple-choice questions, with four alternatives for each item. Item stems and distractors were in English or Pinyin romanization, and drew on Pinyin, character or diagrammatic stimulus material presented in a separate magazine

(Appendix 2). The textual stimulus materials took the form of:

(a) an extract from what might be a letter to a pen-friend, in which the writer introduces herself and her family;
(b) a page from an appointments diary, or calendar;
(c) a brief conversation about the weather and swimming;
(d) an advertisement for a baby-sitter, and details of several applicants for the position;
(e) a conversation between a shop assistant and a customer;
(f) a brief three-way exchange concerning the way to the library.

Table 4 sets out the script-types and stimulus subject-matter of each item. See Appendix 3 for an English translation of the paper.

Item nos	Question	Stimulus	Subject
1-2	Pinyin	Pinyin	Self-introduction
3	English	Pinyin	Self-introduction
4-7	Pinyin	Illustration	n/a
8-10	Pinyin	Character	Diary
11	English	Character	Diary
12-15	English	Character	Weather
16-19	English	Character	Baby-sitting
20-23	Pinyin	Pinyin	Shopping
24-25	English	Pinyin	Directions

Table 4 *Summary of test items*

Topics, functions and vocabulary were selected from those covered in the textbook *Hanyu*, which at that time was almost always used as the major text for beginning students of Chinese in Australia and New Zealand. Favourable teacher feedback confirmed the suitability of the tasks.

Items were scored dichotomously, and the results were analysed by the Australian Council for Educational Research, using the program MSCALE, version 2.4 (Wright, Congdon and Rossner 1987). Items were grouped into three bands, according to their level of difficulty.

Descriptors for each band were drawn up in non-technical language to outline student achievement within each of the three score-ranges (see table 5). Every participating student received a certificate which included the descriptor appropriate to the level at which he or she had performed.

Level 1 (Questions, starting with the easiest, were 3, 1, 4, 2, 24, 13, 25.)
A typical student at this level can read Pinyin and understand some everyday words and phrases, such as names of family members, numbers and time, and pick out simple information from a short piece of writing or dialogue.
Level 2 (Questions were 20, 12, 14, 7, 18, 10, 6, 8, 17, 23.)
A student at this level can read Pinyin and some Chinese characters covering everyday words and phrases such as time, dates and the location of people and objects, and pick out a piece of information from a dialogue or advertisement.
Level 3 (Questions were 21, 16, 5, 9, 19, 15, 22, 11.)
A typical student at this level can understand a range of everyday words and expressions written in Pinyin or Chinese characters, pick out information from a piece of writing and make inferences from the information given.

Table 5 Reading test level descriptors (Source: ACER 1990)

Rasch output

Rasch analysis places the difficulty of items and the ability of individuals on a common scale, and is predicated on the idea that higher ability implies a higher probability of correct responses to items on the scale (Wright and Stone 1979: 44). Part of the output takes a form similar to Table 6. Difficulty of items and ability of candidates are each expressed in units called 'logits', the average difficulty being set at 0.0 logits. The hardest items appear on the top left-hand side, and the easiest on the bottom left-hand side. Similarly, performances of

Logits	persons	items
5.0		
4.0		
	XXXXXXXX	
	X	
3.0		
	XXXXXXX	
	X	
	XXXXXXXX	
2.0	XXXXX	
	X	
	XXXXXXXXXX	
	XXXXXXXXX	
	XXXXXXXXXX	11
1.0	XXXXXXXXX	15 19 22
	XXXXXXXXXXX	5 9 16 21
	XXXXXXXXXX	23
	XXXXXXXXXXX	8 17
	XXXXXXXX	6 10
0.0	XXXXXXXXX	7 18
	XXXXXXX	12 14
	XXXXXXXX	20
	XXXXX	25
	XXXXX	13
-1.0	XX	2 24
	XX	
	X	1 4
-2.0		3
	X	
-3.0		

Each 'X' in the person column represents 4 persons.

Table 6 *Map of persons and items*

candidates are represented from highest (top) to lowest (bottom) on the right-hand side.

Candidates demonstrating an ability equivalent to any given point on the scale would have a 70% (in this case)[3] probability of supplying the correct response to items at the same level. They would generally be expected to answer correctly items which appear below that point on the scale, and answer incorrectly those above it (although departures from the pattern are expected).

Test results

Of the 592 candidates the results of 551 were included in the IRT analysis. The remaining candidates answered either all or none of the 25 items correctly, and could therefore not be included in the analysis. No ability estimates could be produced for these candidates because the test was inappropriate for them.

An important step in Item Response Modelling is the estimation of goodness of fit. The index used here is the weighted mean square, the ratio of the spread of responses to the expected spread. The expected value is 1.0, but items with mean square values between .8 and 1.2 are assumed to be functioning well and measuring the same trait as the others. All of the 25 items fell within the acceptable range. A similar test is applied to person fit, but since this paper is concerned with the performance of candidates as a group, person fit will not be discussed here. Table 7 shows the difficulty and fit of each item.

METHODOLOGY

At this point we will attempt to establish the extent to which the factors listed in Table 2, as well as test-method factors, were related to item difficulty and hence to the performance of candidates on the test.

For the test items, key phrases have been isolated (table 8). These phrases may be drawn from the question or the stimulus material or both, and are central to the understanding of the items in which they occur. They are the part(s) of the stimulus text in which the answer is located, and where appropriate, the parts of the test item which pose the problem and express the key (correct alternative).

Item No.	Logits	Weighted MnSq (infit)
1	-1.62	0.97
2	-0.94	1.03
3	-1.95	1.10
4	-1.52	1.00
5	0.73	1.06
6	0.19	0.89
7	0.08	0.98
8	0.39	0.88
9	0.77	0.99
10	0.20	0.93
11	1.19	1.02
12	-0.22	0.93
13	-0.76	0.96
14	-0.19	0.97
15	0.99	1.18
16	0.74	1.15
17	0.49	0.99
18	0.07	0.95
19	0.95	0.91
20	-0.43	0.96
21	0.75	1.11
22	1.07	0.98
23	0.60	1.08
24	-0.91	1.03
25	-0.67	0.99

Table 7 Item parameter estimates

Three points should be made about the key phrases. Firstly, selection of the key may, in some instances or for some students, be more a process of elimination than recognition of the correct alternative. Lumley, in his study based on a post-hoc analysis of a reading test of English as a Second Language, points out, 'one cannot judge what

happens inside students' heads' (1993: 46). This is a perennial problem in this kind of investigation, and indeed for the writers of the band descriptors used in the kind of reporting exemplified here. However, Lumley goes on to say that despite this, it is demonstrably possible for investigators (in this case teachers):

> to agree in their speculation about the processes involved in answering test items. This speculation may be informed by some intuition about what makes an item hard, or what differentiates it from other items...the sub-skills identified by the teachers are not claimed to be the only things tested by the test items, but rather an expression of what they considered to be skills without which the question could not be answered.
>
> (1993: 47)

The second point to be made about the key phrases refers to the fact that the literature surveyed in the early part of this paper dealt largely with recognition of words in isolation. However, for several of the statistical tests conducted in this study, the characteristics of individual words are averaged across the key phrase. So, to take word length and item 1 as an example, the number of morphemes in the phrase is four, and this number, when divided by the number of words in the phrase, gives a word length figure of 1.33. It is acknowledged that this approach may mute or blur the effect of individual words. It was taken because, in an attempt to make this test as communicative as its format allows, the test designers have combined vocabulary in various ways; this is not a test of vocabulary in isolation.

Finally, there is a certain artificiality in this isolation of key phrases. The meaning of each phrase of course depends to some extent on the surrounding text. However, for reasons argued earlier, word knowledge is seen as crucial in elementary-level reading, and even more so in the context of this test.

The key phrases were examined for the presence of the factors listed in Table 3 above. The procedure in each case is described under the relevant heading below. A matrix of information derived from this process can be found in Appendix 4.

Pearson's *r* correlations were used to test the relationship between item difficulty and those factors (Word Length, Character Complexity and Order of Presentation) which were present to varying degrees in different items. For the remaining factors, t-tests were used. In some

Item	Phrases (Chinese)	Phrases (English)
1	Ta you didi	She has a little brother
2	dou shi laoshi; ye	are both teachers; too
3*	shisi sui	fourteen years old
4	Qiche shangmian you shenme?	What is on the car?
		mao a cat
5	Qiu zai nali?	Where is the ball?
	qiche qianmian	in front of the car
6	Qiche houmian you jige ren?	How many people are there behind the car?
	yige	one
7	Qiche shangmian you meiyou qiu?	Is there a ball on the car?
8	Jintian shi xingqiji?	What day is today?
	xingqisi	Thursday
9	Zuotian shi xingqisan.	Yesterday was Wednesday.
10	Mingtian shi xingqiwu.	Tomorrow is Friday.
11 *	*si dian zu qiu*	*football four o'clock*
12*	*re*	*hot*
13*	*kuai shiyi*	*nearly eleven*
14*	*women qu youyung*	*Let's go swimming*
15*	*Xianzai tai mang, xiawu qu*	*I am too busy now; Let's go this afternoon.*
16*	*nu xuexeng*	*female student*
17*	*xihuan xiao haizi*	*likes small children*
18*	*xihuan xiao haizi*	*likes small children*
19*	*shuo hanyu*	*speaks Chinese*
20	yao mai shenme?	What (do you) want to buy?
	you meiyou	Have you any
21	gei shouhuoyuan	gives the shopkeeper
	liangkuai qian	two dollars
22	shouhuoyuan zhao gei;	The shopkeeper gives change;
	liangmao qian	20 cents
23	hai yao mai Ying-Han cidian	also wants to buy an English-Chinese dictionary
24*	Wo bu zhidao	I don't know
25*	Zai youju pangbian	Beside the Post Office

Underlined phrases are drawn from the stimulus text only.
*Stems and alternatives of these items were presented in English.
Italicised phrases appear in characters in the test paper.

Table 8 Key phrases

cases this was because of the nature of the factors themselves: they represent a dichotomy (Chinese/English, Pinyin/characters, etc). Other factors, were treated as though a dichotomy existed (e.g. whether or not the item contained unknown vocabulary) because they appeared too rarely to offer the kind of gradation necessary for a correlation. All t-tests were one-tailed, the hypothesis being that item difficulty would increase in the presence of any of the following: pronunciation difficulty, unfamiliar syntactic patterns, multiple meaning, Chinese language questions, character-script stimulus materials, unknown words and lexical items which had been presented in the same unit as closely related words and phrases.

Word-form factors: word length and pronounceability

For word length, the mean number of morphemes per word was calculated, with word boundaries being established as outlined earlier in this paper. For Pinyin script, word boundaries were taken to be those observed by the test paper. For character script, word boundaries were taken to correspond to those observed by the textbook (Chang, Mackerras and Yu 1990). For pronounceability the presence of 'problem' lexical items, viz those with the initial consonants *q* (with no following *u*) *zh*, and *x*, was noted.

Grammatical characteristics : derivational complexity and syntactic patterns

Because word-length (above) was calculated in terms of morphemes, this aspect of derivational complexity has already been covered. The second aspect, transparency of complex words, did not lend itself to investigation because almost every key phrase contained one or two polysyllabic words whose meaning was not transparent. The presence of syntactic patterns unlike those of English was noted. Examples of unfamiliar syntactic patterns would include stative verbs such as *mang* 'busy' whose approximate equivalents in English are adjectives, postpositions such as *pangbian*, which are translated by English prepositions, and the use of subject-verb-object word order with interrogatives such as *shenme* 'what'.

Semantic features: codability

Key phrases were examined for the presence of lexical items with no close equivalent in English. Because only five of these were found (perhaps unsurprisingly, since this is an elementary-level test), no further tests for the influence of codability were carried out. However, as table 9 shows, the impression given by the small sample coincides with the expectation that low codability is associated with item difficulty. All of the items containing words and phrases with low codability in English showed above average difficulty (the average difficulty of the test was 0.00 logits). No cognates were found in key phrases.

Item	Word	Logits
5	zai	0.73
6	ge	0.19
8	xingqiji	0.39
21	liangkuai	0.75
22	zhaogei	1.07

Table 9 Codability and item difficulty

Multiple Meaning: polysemy and homonymy

An example of multiple meaning was considered to exist if any lexical item in the key phrase had another meaning which was presented in *Hanyu* (Chang, Mackerras and Yu, 1990). These are listed in table 10. Homonyms are usually represented by different characters. However, because *Hanyu* presents most vocabulary in Pinyin form before characters, and most often recycles vocabulary in Pinyin form, multiple meaning was taken to be equally relevant for both script-types in the test. For the purposes of establishing multiple meaning, word boundaries were observed in the same manner as for Word Length (above).

Written form: script and character complexity

The script (characters or Pinyin) used for the stimulus text, and the language of the item were noted. The mean number of strokes per character was calculated for each key phrase.

Item	Word	Meaning	Alternative meaning
2	shi	are	business
5	zai	is (at)	again
8	shi	is	business
9	shi	is	business
10	shi	is	business
11	dian	o'clock	a little, dot
13	kuai	nearly	fast
15	tai	too	Mrs
22	qian	money	front
22	zhao	give change	look for, find
25	zai	is (at)	again

Table 10 Multiple Meaning

Extralingual factors: order of presentation, unknown lexical items, grouping

The textbook *Hanyu* (Chang, Mackerras and Yu 1990) used by most of the candidates presents new vocabulary gradually over eight units (the final unit is for revision only). Once presented, words and phrases are continually recycled in later units. Of the 27 schools responding to the questionnaire, 14 indicated that they had completed the first eight units, and eleven had completed the first seven units. The effect of the fact that some schools did not complete Unit 8 was minimised by the fact that only five items (numbers 16, 20, 21, 22 and 23) included any vocabulary from Unit 8. The design of the textbook seems to assume that the units will be covered sequentially, and examination of information derived from the teachers' questionnaires suggests this is usually the case. The assumption was then made that there is an approximate correspondence between the order of presentation of vocabulary in the textbook and students' familiarity with that vocabulary. The number of the unit (1-8) in which each lexical item first occurs was noted, and this number used as a value representing the position of the vocabulary item in the sequence

of presentation. Lexical items which do not appear in the textbook (unknown words) were allocated a value of 9. This conservative value simply takes account of the fact that a word or phrase which has not been presented be will be less familiar than one that has, however recently. A mean was calculated for each key phrase to indicate the relative familiarity of the phrase as a whole.

The presence in key phrases of words which are not introduced in the textbook was noted. Finally, each key phrase was examined for the presence of lexical items which were presented in the same unit as other words and phrases from the same lexical field.

Method effects

The multiple-choice format of the ALC test would probably not account in any way for differences in item difficulty, since all of the items on the test are of this type, but it could mean that the trait which is measured is to some extent not the one intended.

We will now consider differences between task-types within the multiple-choice format. As Table 11 shows, each of the test items required one or more of the following activities: locating information within a text, translating, matching, combining information. Locating information, present in items types i, iii and vii, was the most frequently required activity.

	Item type	Item Number
i	locate and translate/ rephrase	11 12 20
ii	match	1 8 21 22 23
iii	locate and combine	16 17 18 19
iv	diagram match	4 5 6 7
v	translate	3 13 14 24 25
vi	combine	2 15
vii	locate	9 10

Table 11 Task types

There has been some debate on the question of whether subskills can be accurately identified and whether a hierarchy of reading subskills exists at all (see for example Alderson and Lukmani, 1989; and Lumley, 1993). Lumley has shown that reading sub-skills are 'a useful construct with which...test constructors may work' (1993: 50), and that, with some care, agreement can be reached about which subskills are being tested in reading test items. It must be admitted, however, that the kind of care taken with the descriptors in Lumley's study requires more time than is sometimes available for this part of the ALC process.

The effects of some of the task-types are not open to investigation here because there is a correspondence between certain task-types (particularly 'locate and combine' and 'diagram match') and input material. The effects of the task-types are likely to be confounded with other features of the input text.

According to Bachman (1990: 128) test performance can be affected by the extent to which the test item directs candidates' attention to a specified and limited amount of material. If this kind of direction has influenced item difficulty in the ALC test, we might expect to see the difficulty levels of items which require candidates to locate information (i.e. items which do not focus candidates' attention on a small section of text) to be more difficult than those that do not.

RESULTS

The results of the correlations are shown in table 12 below.

	Count	Covariance	Correlation	R^2
Word length	25	.083	.319	.102
Character complexity	12	.161	.151	.023
Order of presentation	25	.823	.524	.274

Table 12 Correlation between item difficulty and word length, character complexity and order of difficulty

The results for the t-tests are shown in table 13. The results were significant at .05% for pronounceability, the script used in the stimulus material, the presence of unknown words and the order of presentation.

Group	Count	Mean	Std. Dev.	Std. Error	DF	t-value	Prob
Pronounceability problems present	14	.39	.692	.185			
No pronounceability problems	11	-.496	.905	.273			
Significance					23	2.778	.0054
Unfamiliar syntactic patterns present	14	.102	.848	.227			
No unfamiliar syntactic patterns	11	-.13	.979	.295			
Significance					3	.635	.2658
Multiple meaning present	10	.297	.809	.256			
No multiple meanings	15	-.198	.922	.238			
Significance					23	.1379	.0907
Pinyin text stimulus	9	-.456	1.059	.353			
Character script stimulus	12	.385	.585	.169			
Significance					19	-2.329	.0155
Pinyin text questions	13	.021	.886	.246			
English questions	12	-.023	.945	.273			
Significance					23	.118	.4535
'Unknown' words present	11	.35	.684	.206			
No 'unknown' words present	14	-.275	.968	.259			
Significance					23	1.812	.0415
Vocabulary presented in lexical sets	11	-.095	.987	.297			
Vocabulary not presented in lexical sets	14	.075	.848	.227			
Significance					23	-.464	.3233
Locate task-type	9	.418	.548	.183			
Other task-type	12	-.27	1.024	.295			
Significance					19	1.821	.0422

Table 13 *Relationship between item difficulty and pronounceability, syntactic complexity, multiple meaning, script-style, unfamiliar vocabulary and grouping for presentation*

DISCUSSION

The analysis above indicates that a significant and positive relationship was found between item difficulty and the presence of unknown words, the candidate's unfamiliarity with words in the test item (as measured by the order of presentation), the use of characters rather than Pinyin in the stimulus text, and the perceived difficulty of pronouncing words included in the item. The test for multiple meaning also approaches significance.

As table 11 shows, the 'locate' task type was associated with a higher level of difficulty than items which did not require candidates to locate information. It is not clear whether the task-type is purely a test-method facet, or whether it has a linguistic component.

The analysis so far has been in terms of key phrases. As table 7 indicates, these were almost invariably located in the question stem and alternatives of items written in Pinyin, but for items presented in English, the phrases were, of course, found in the stimulus material. We will now turn our attention to the input texts on which the test items draw.

In table 14 the stimulus texts are arranged from hardest (at top) to easiest, according to the mean difficulty (in logits) of items associated with them. The table also shows the mean score for the order of presentation (O-P) of the lexical items in each text, the total number of unknown words (UW) in each, and the script-style in which the text was presented. It was not possible to quantify pronounceability in a way that would allow meaningful comparisons between stimulus texts. Items 4-7 have been omitted because they relate to a diagram, rather than a text, and the only reading which is tested in them is the reading of the items themselves.

Logits	Stimulus	Items	O-P	UW	Script
0.64	Calendar	8-11	6.79	6	Chars
0.50	Shopping	20-23	5.85	13	Pinyin
0.56	Conversation 1	16-19	4.74	1	Chars
-0.50	Babysitting	12-15	4.00	0	Chars
-0.79	Conversation 2	24-25	4.30	1	Pinyin
-1.50	Letter	1-3	2.82	0	Pinyin

Table 14 *Input text characteristics, in difficulty order*

The information in table 14 generally supports our earlier findings: texts associated with the highest level of item difficulty are the three character texts, and the Pinyin text which contained the greatest number of unknown words.

Except for the 'Babysitting' text, the order of presentation index increases with item difficulty. The difficulty of this text may have been increased by the nature of the tasks which were attached to this text. These items required candidates to combine information from different parts of the text, a skill not required by items drawing on the remaining texts. The skill of combining information from different parts of a text was not identified in the statements describing the three skill levels (see the discussion of descriptors, below).

While differences between texts can be distinguished, it should not be assumed that a stimulus text will necessarily be associated with test items of more or less uniform difficulty. Item difficulty can have its source in the question itself, the stimulus material, or both. It may be linguistic in origin, or a method effect. Consider, for example, item 15. The item draws on the conversation written in character text, but is itself presented in English:

15. Chen Xiaolan...

A	says the time is passing slowly.
*B	can't go until later.
C	suggests he go with someone else.
D	suggests they go out tonight.

This item had a difficulty index of 0.99, and was the third most difficult item on the paper. The other items (numbers 12, 13 and 14) which related to the same input text had difficulty levels of -.22. -.76 and -.19, and were ranked in places 22, 20 and 16 respectively.

Item 15 had a facility of 49%, while its most commonly chosen distractor (C) attracted 32% of the candidates. Candidates who chose alternative D did in fact correctly identify the function and topic of Chen Xiaolan's second turn in the conversation, but apparently mistook the meaning of *xiawu* 'afternoon' for 'evening'. Candidates who chose the key correctly conflated Chen Xiaolan's two statements (*I am too busy now; Let's go this afternoon*) into one which interpreted her words, rather than merely attempting to translate them. This is rather more demanding

than most of the other items on the paper, and the difficulty is linguistic in nature.

As an example of difficulty which is not necessarily linguistic let us consider items 17 and 18.

17. Dawei is suitable for the job because he...

A	loves cats.
B	speaks English.
C	is a high school student.
*D	loves children.

18. Mali is suitable for the job because she...

*A	loves children
B	is the right age.
C	is a teacher.
D	speaks English.

The two items relate to the same stimulus text. Their difficulty levels are 0.49 and 0.07 respectively, and they ranked tenth and fifteenth in difficulty. Since the stems of the two questions are almost the same, and the keys are identical, the difference in difficulty levels presumably has its source in the distractors. The lesser difficulty level of item number 18 appears not to be linguistic, but an unidentified method effect. It might, for instance be a context effect (Wood 1991: 15): i.e. the result of its positioning immediately after a similar item.

Bachman reports that 'numerous research studies...have demonstrated that the methods we use to measure language ability influence the performance on language tests' (1990: 113). It is not possible to avoid these effects completely; our task as testers is to acknowledge and understand them, and make our tests valid despite them.

DESCRIPTORS

We will now turn to the descriptors (table 5) which were drawn up to describe performances within score ranges. The aim here is to see how far the descriptors correspond with the dimension constructed by the Rasch scaling, and with the reading-skill factors which have been shown to be related to that scale.

Logits	persons	items	
5.0			
4.0			
	XXXXXXXX		
	X		
3.0			
	XXXXXXX		
	X		
	XXXXXXXX		
2.0	XXXXX		
	X		BAND 3
	XXXXXXXXXX		
	XXXXXXXXX		
	XXXXXXXXX	11	
1.0	XXXXXXXXX	15 19 22	
	XXXXXXXXXXX	5 9 16 21	
	XXXXXXXXXX	23	
	XXXXXXXXXX	8 17	
	XXXXXXXX	6 10	
0.0	XXXXXXXXX	7 18	BAND 2
	XXXXXXX	12 14	
	XXXXXXXX	20	
	XXXXX	25	
	XXXXX	13	
-1.0	XX	2 24	
			BAND 1
	XX		
	X	1 4	
-2.0		3	
	X		
-3.0			
-4.0			
-5.0			

Each 'X' in the person column represents 4 persons.

Table 15 *Map of persons and items, showing band cut-offs*

The scores of the participants were mapped onto the same scale as the one represented in the map of persons and items (table 6). The scores and items were then grouped into three bands (Band 3 being the highest, and Band 1 the lowest), and a descriptor for each band was formulated. Table 15 shows the application of these bands to the scale. The cut-off points for the bands were necessarily arbitrary, and in this case were based principally on the clustering of candidates' scores. The placement of the cut-offs was influenced by the purpose of the test (to encourage language learners); that is to say, the cut-off points were positioned in a way that ensured that a high proportion of the participants was placed in the higher bands. It is probable, therefore, that the cut-off points do not represent discrete sub-skills of reading, but can be interpreted as approximations or general descriptions of tendencies in the developing repertoire of foreign language reading skills of the candidate.

The descriptors were based on a post-hoc analysis of the content of the test items falling within each of the three bands. The writers of the descriptors were constrained by:

(a) the need to make statements intelligible to both school children and their parents...
(b) the limited space on the certificate...
(c) the desire to give all learners a sense of achievement...
(d) the need for equivalence across languages (Elder and Zammit 1992: 15).

Each year the process of writing the descriptors begins only after statistical analysis of the students' performances is complete, and must be finished in time for the certificates to be printed and distributed. Obviously, the intended effect of encouragement to students will be lost if the process is not completed quickly. The pressure of time no doubt has some effect on the thoroughness with which this part of the task is carried out, and does not allow for the kind of discussion which may be necessary to ensure that the descriptors amount to more than the interpretation of one or two people.

The descriptors for 1990 name the rather broadly conceptualised abilities to read Pinyin and characters as well as three other major sub-skills: recognising vocabulary within certain semantic fields, picking out information and making inferences.

The first of these, vocabulary recognition, appears to coincide with our findings regarding order of presentation. At the Level 1, two of the three semantic fields mentioned are dealt with in Unit 3 of the text-book. The third, telling the time, occurs in Unit 7, but closer examination of the item in question (item 13) shows that to answer it correctly the candidates needed only to recognize the number 'eleven'. Numbers are introduced in Unit 3. At level 2 the semantic fields identified are presented in units 6 and 7. The descriptor for Level 3 does not identify the semantic fields needed for this level.

The second sub-skill noted in the descriptors, picking out information, is probably the equivalent of our 'locating of information' discussed above. This sub-skill or task appears in items 9, 10, 11, 12, 17, 18, 19 and 20, i.e. in levels 2 and 3 only. The description is therefore meaningful as a way of distinguishing Level 1 from Level 2, and supports the suggestion above, that the tendency of some items to focus candidates' attention on a limited section of the text will affect item difficulty. Making inferences is generally considered to be a higher level skill, and it might be argued that the term 'inference' is a somewhat inflated description of the processes required by items on this paper.

The cut-off points between clusters of item types are arbitrary, and the grouping of sub-skills within the levels is not clear-cut. The task of drawing up meaningful descriptions of the levels is therefore a difficult one. Nonetheless, it can be said that the descriptors above go some way towards defining the three ability levels, and do not appear to have been unduly prejudiced by the effects of unexamined test facets. The dimension mapped out by the IRT scaling can be explained principally in terms of relative familiarity and orthographic characteristics of the stimulus text.

CONCLUSION

This paper has examined the relationship of predictive variables to difficulty levels of foreign language reading test items. It has shown that certain of these predictors, in combination with aspects of the task itself, are related to the item-difficulty order established by the Rasch scaling procedure. This relationship gives support to the practice of using Rasch scaling as a basis for descriptions of candidates' performance on a test of the type described in this study. The descriptors themselves have been

shown to have provided an adequate, if incomplete, description of the foreign language reading skills demonstrated by candidates at various levels of performance. Constraints of the kind faced by the writers of the ALC descriptors would be unlikely to apply in the case of high-stakes tests. With greater freedom at the reporting stage, and more detail in the descriptors, content-referenced interpretations of test scores should be able to reflect even more faithfully the being skills assessed.

The methodology described in this paper could be used as a way of investigating ALC content each year, and for each language, ideally in time to provide information on which better descriptors could be developed. The information gathered from the process could inform the developers of new test material for the programme, and provide a check on the extent the match between test specifications and content. Similarly, the methodology might be usefully applied elsewhere for these kinds of purposes. The greater the consequences of the test, the more important it is to understand its content, and to ensure the accuracy of its descriptions of candidates' performances.

NOTES

1 The Language Certificates are an initiative of the Australian Multicultural Foundation.

2 Pinyin is the most widespread of several systems for writing Mandarin Chinese using the Roman alphabet. Pinyin is the romanization system officially recognized by the People's Republic of China, and it is the one used by the textbook *Hanyu*, as well as the ALC test papers.

3 Items are usually located on the same scale at the 50% probability level, but for some purposes (as for this test), the level may be varied.

REFERENCES

Alderson, J.C. and Y. Lukmani (1989) Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language* 5,2: 253-270.

Alderson, J.C. and A.H. Urquhart (1984) Reading in a *Foreign Language*. London, Longman.

Allen, E.D. and R.M. Vallette (1972) *Modern Language Classroom Techniques: a Handbook*. New York, Harcourt Brace Jovanovich.

Anderson R. and P. Freebody (1981) Vocabulary knowledge. In J.T. Guthrie (ed.) *Comprehension and Teaching: Research Reviews*. 77-117. Delaware, IRA.

Australian Council for Educational Research (1989) *Basic Skills Testing Program: Technical Report*. Melbourne, ACER.

Australian Council for Educational Research (1990) *Report on the Australian Language Certificates to the Australian Bicentennial Multicultural Foundation*. Melbourne, ACER.

Bachman, L.F. (1990) *Fundamental Considerations in Language Testing*. Oxford, OUP.

Ballas, M. (1987) National assessment broadens mission to provide answers for parents, educators, policymakers. *ETS Developments* 32,4: 5-7.

Blum, S. and E. A. Levenston (1978) Universals in lexical simplification. *Language Learning* 28:399-416

Chall, J.S. (1987) Two vocabularies for reading: recognition and meaning. In M.G. McKeown and M.E. Curtis (eds) *The Nature of Vocabulary Acquisition*. Hillsdale, N.J., Lawrence Erlbaum and Associates.

Chang, P. A. Mackerras and H-C Yu (1990) *Hanyu 1: Chinese for Beginners*. Melbourne, Longman.

Cziko, G.A. (1980) Language competence and reading strategies: a comparison of first- and second-language strategies. *Language Learning* 30,1: 100-116.

Dale, E. and J.S. Chall (1948) A formula for predicting readability. *Educational Research Bulletin*. 27: 11-28.

Davis, F. (1944) Fundamental factors of comprehension in reading. *Psychometrika* 9,3: 185-197.

Dollerup, C. E. Glahn and C.R. Hansen. (1989) Vocabularies in the reading process. In P. Nation and R. Carter (eds) *Vocabulary Acquisition*. Amsterdam, AILA.

Dubin, F. (1989) The odd couple: reading and vocabulary. *English Language Teaching Journal* 43,4: 283-287.

Elder, C. and S. Zammit (1992) Assessing performance in languages other than English: the contribution of the Australian Language Certificates. *Vox*. 6: 14-20.

Everson, M. F. (1988) Speed and comprehension in reading Chinese romanization vs characters revisited. *Journal of the Chinese Language Teachers' Association* 23,2: 1-15.

Flesch, R.F. (1948) A new readability yardstick. *Journal of Applied Psychology*. 32:3 221-233.

Freebody, P. and R Anderson (1983) Effects of vocabulary difficulty, text-cohesion and schema availability on reading comprehension. *Reading Research Quarterly* 18,3: 277-294.

Fry, E. (1968) A readability formula that saves time. *Journal of Reading* 11: 513-516.

Gibson, E.J. and H. Levin (1975) *The Psychology of Reading*. Cambridge, Mass, The MIT Press.

Goodman, K.S. (1976) Reading: a psycholinguistic guessing game. In H. Singer and R. Ruddell (eds) *Theoretical Models and Processes of Reading*. Newark, DE, International Reading Association.

Hague S.A. (1987) Vocabulary instruction: what L2 can learn from L1. *Foreign Language Annals* 20,3: 217-225.

Halliday, M.A.K., A. McIntosh and P. Strevens (1964) *The linguistic science and language teaching*. London, Longman

Hatch E. (1983) *Psycholinguistics: a Second Language Perspective*. Rowley, Mass, Newbury House.

Hatch, E. (1974) Research on reading in a Second Language. *Journal of Reading Behaviour* 6,1: 53-61.

Hayes, E.B. (1987) The Relationship between Chinese character complexity and character recognition. *Journal of the Chinese Language Teachers' Association* 22,2: 45-72.

Henning, G.H. (1973) Remembering foreign language vocabulary: acoustic and semantic parameters. *Language Learning* 23: 185-196.

Higa, M. (1965)The psycholinguistic concept of 'difficulty' and the teaching of foreign language vocabulary. *Language Learning* 15,3&4: 167-179.

Huckin, R., M. Haynes and J. Coady (eds) (1993) *Second Language Rreading and Vocabulary Learning* . Elwood NJ, Ablex

Juel, C. (1983) The development and use of mediated word identification. *Reading Research Quarterly* 18,3: 306-327.

Koda, K. (1994) Second language reading research: Problems and possibilities. *Applied Psycholinguistics* 15: 1-28.

LaBerge, D. and S.J. Samuels. (1974) Toward a theory of automatic information processing in reading. *Cognitive Psychology* 6,2: 293-323.

Laroche J.M. (1979) Readability measurement for foreign language materials. *System* 7,2: 131-5.

Laufer, B. (1989) A factor of difficulty in vocabulary learning: deceptive transparency. In Nation, P. and R. Carter (eds) *Vocabulary Acquisition. AILA Review 6*: 10-20. Amsterdam, AILA.

Laufer, B. (1990) Why are some words more difficult than others? Some intralexical factors that affect the learning of words. *IRAL* 28,4: 293-307.

Ludwig, J. (1984) Vocabulary acquisition as a function of word characteristics. *The Canadian Modern Language Review* 40: 552-562.

Lumley, T.J.N. (1993) Reading comprehension sub-skills: teacher perceptions of content in an EAP test. *Melbourne Papers in Applied Linguistics* 2,1: 25-55.

McKeown, M.G. and M.E. Curtis. (1987) *The Nature of Vocabulary Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

McNamara, T.F. (1990) *Assessing the Second Language Proficiency of Health Professionals*. Unpublished PhD thesis, University of Melbourne.

MacWhinney, B. (1995) Language-specific prediction in foreign language learning. *Language Testing* 12,3: 293-320.

Meara, P. (compiler and ed.) (1983) *Vocabulary in a Second Language*, Volume 1. London: CILT.

Meara, P. (compiler and ed.) (1987) *Vocabulary in a Second Language*, Volume 2. London, CILT.

Meara, P. (compiler and ed.) (1993) Vocabulary in a Second Language, Volume 3. *Reading in Second Language* 9,1: 761-837.

Mossensen, L., P. Hill and G. Masters (1987) *Test of Reading Comprehension (TORCH)* Melbourne, Australian Council for Educational Research.

Mullis, I.V.S. and L.B. Jenkins (1988) *The Science Report Card: Elements of Risk and Recovery*. Princeton, N.J., Educational Testing Service.

Nation, I.S.P. (1990) *Teaching and Learning Vocabulary*. NY, Newbury House.

Palmberg, R. (1987a) On lexical inferencing and the young foreign language learner. *System* 15,1 69-76.

Palmberg, R. (1987b) Patterns of vocabulary development in foreign language learners. *Studies in Second Language Acquisition* 9,2: 201-219.

Pavlidis M. (1990) *Reading Strategies of Second-Language Learners of Chinese*. Unpublished MA thesis, University of Melbourne.

Pino-Silva, J. (1993) Untutored vocabulary acquisition and L2 reading ability. *Reading in a Foreign Language* 9,2: 845-857.

Raugh, M.R. and R Atkinson (1975) A mnemonic method for learning second language vocabulary. *Journal of Educational Psychology* 67:1 16.

Read, J. (1988) Measuring the vocabulary knowledge of second language learners. *RELC Journal* 19,2: 12-25.

Richards, J (1974) Word lists: problems and prospects. *RELC Journal* 5,2: 69-84.

Rumelhart, D.E. (1977) Towards an interactive model of reading. In S. Dornic (ed.) *Attention and Performance* VI. Hillsdale, N.J., Lawrence Erlbaum Associates.

Rumelhart, D.E. (1980) Schemata: the building blocks of cognition. In R.J. Spiro, B Bruce and W.E. Brewer (eds) *Theoretical Issues in Reading Comprehension*, 33-58. Hillsdale, N.J., Lawrence Erlbaum Associates.

Schneider, W. and R.M. Shiffrin (1977) Controlled and automatic human information processing I: detection, search and attention. *Psychological Review* 84,1: 1-66.

Shiffrin R.M. and W. Schneider (1977) Controlled and automatic human information processing II: Perceptual learning, automatic attending, and a general theory. *Psychological Review* 84,2: 127-190.

Stanovich K.E. (1980) Towards an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly* 16,1: 32-71.

Tinkham, T. (1993) The effect of semantic clustering on the learning of second language vocabulary. *System* 21, 3: 371-380.

Williams, E. and C. Moran (1989) Reading in a foreign language at intermediate and advanced levels with particular reference to English. *Language Teaching* 23, 2 : 217-228.

Wood, R. (1991) *Assessment and Testing: A Survey of Research Commissioned by the University of Cambridge Local Examinations Syndicate*. Cambridge: Cambridge University Press.

Wright, B.D., R.T. Congdon and M. Rossner. (1987) *MSCALE, Version 2.4*. University of Chicago.

APPENDIX 1: AUSTRALIAN LANGUAGE CERTIFICATE TEST PAPER

READING COMPREHENSION

The following tasks are to see how well you can understanding what you read in Chinese. First read the article or story in the magazine, then read the questions in the question booklet and finally mark your answer on the answer sheet. Your score is based on answers correctly marked on your answer sheet. Do NOT write anything on the magazine or question booklet.

For each question you are to choose the ONE answer that you think is best.

Example:

Look at the clock above and fill in the blank with the correct time.

Xiànzài ..... diǎnle.
A liù B sān C sì D shí

The correct answer is D. *As the letter* D *represents this answer, record your answer on the answer sheet like this:*

A ☐ B ☐ C ☐ D ▬

Find the photograph of Liú Xiǎohóng on Page 1 and read what she has to say. From the information she gives, mark the answer which would correctly complete the sentences about her.

1. Tā yǒu .....
   A dìdi
   B mèimei
   C jiějie
   D gēge

2. Tā māma hé bàba .....
   A bù dōu shì lǎoshī
   B bù yě shì
   C dōu shì lǎoshī
   D dōu búshì lǎoshī

3. She is .....
   A 10 years old
   B 40 years old
   C 11 years old
   D 14 years old

Look at the picture of the car on Page 1 of your magazine. Mark the correct answer to each of the questions given below.

4. Qìchē shàngmian yǒu shénme?
   A yǒu rén
   B yǒu māo
   C méiyǒu shénme
   D yǒu gǒu

5. Qiú zài nǎli?
   A zài qìchē hòumian
   B búdùi
   C zài qìchē qiánmian
   D zài gǒu xiàmian

6. Qìchē hòumian yǒu jǐge rén?
   A jǐge
   B yǒu gǒu
   C yǐge
   D sānge

7. Qìchē shàngmian yǒu méiyǒu qiú?
   A yǒu sānge
   B yǒu
   C yǒu rén
   D méiyǒu

On page 2 you will find a page from Wáng Píng's diary. Pretend you are Wáng Píng and answer the following questions.

8. Jīntiān shì xīngqīji?
   A xīngqīsì
   B xīngqīsān
   C sì tiān
   D liù diǎnzhōng

9. Zuótiān shì.....
   A 1989 nián
   B qīhào
   C xīngqīsān
   D 1991 nián

10. Míngtiān shì.....
    A yīyuè wǔhào
    B xīngqīwǔ
    C jīntiān
    D 1991 nián

11. Which of the following is true?
    A You will play football at 4.
    B Today is not a school day for you.
    C You will go home at 4 o'clock.
    D You will have lunch at school.

Read the dialogue between Xiǎo Hú and Chén Xiǎolán on Page 3 of the magazine and then mark the item which completes each sentence correctly.

12. The weather is.....
    A rainy
    B cold
    C hot
    D like yesterday's

13. It is.....
    A just after ten o'clock
    B twelve o'clock
    C early afternoon
    D nearly eleven o'clock

14. Xiǎo Hú.....
    A wants to go and listen to music
    B is going swimming with his father
    C suggests the two of them go swimming
    D wants to go and do something useful

15. Chén Xiǎolán.....
    A says the time is passing slowly
    B can't go until later
    C suggests he go with someone else
    D suggests they go out tonight

On Page 3 of your magazine you will find a copy of an advertisement for a babysitter which was in the window of the local milk bar. Read it carefully and then answer the question below:

16. Which of the following might get the job?
    A a 17-year old boy
    B a grade 6 student
    C a female high school student
    D a young high school teacher

There are 4 people who are interested in the job advertised: Dàwèi, Mǎlì, Sū and Annà. Read the information about each one and read the advertisement again. Then complete the following sentences correctly according to the information provided.

17. Dàwèi is suitable for the job because he.....
    A loves cats
    B speaks English
    C is a high school student
    D loves children

18. Mǎlì is suitable for the job because she.....
    A loves children
    B is the right age
    C is a teacher
    D speaks English

19. Both Ānnà and Sū can apply for the job because they.....
    A love children
    B speak Chinese
    C go to the University
    D can work in the mornings

Complete the sentences below according to what you know happened in the Singapore shop according to the story on Page 4 of the magazine.

20. Ālún yào.....
    A mǎi běnzi
    B gěi běnzi
    C kàn běnzi
    D zhǎo běnzi

21. Tā gěi shòuhuòyuán.....
    A liùmáo qián
    B liǎngkuài qián
    C liǎngmáo qián
    D shíkuài qián

22. Shòuhuòyuán zhǎo gěi Ālún.....
    A yíkuài bā
    B liùmáo qián
    C yíkuài qián
    D liǎngmáo qián

23. Ālún hái yào mǎi.....
    A Yīngwén cídiǎn
    B máobǐ
    C Yīng-Hàn cídiǎn
    D Hànzi

Look at the drawing of Anne Miller on the Back Page of the magazine and then answer the following questions.

24. The man says to Anne.....
    A it's over there
    B I don't know
    C I don't understand
    D there isn't one around here

25. The woman says to Anne.....
    A beside the post office
    B I don't understand
    C it's over there
    D post your letter at the hotel

STOP HERE

APPENDIX 2: MAGAZINE

你好!
Wǒ jiào Liú Xiǎohóng. Wǒ zhù zài Shànghǎi. Wǒde jiā yǒu sìge rén: bàba, māma, dìdi hé wǒ. Wǒ bàba shì lǎoshī, wǒ māma yě shì. Wǒ shísì suì. Wǒ shàng zhōngxué. Dìdi shì xiǎoxuéshēng. Tā hěn xǐhuan tī zúqiú. Wǒ bù xǐhuan, wǒ xǐhuan dǎ lánqiú.

BEST COPY AVAILABLE

Tài kěxiào! What a joke!

Qǐng wèn, zhètiáo lù shì dào nǎr qùde?
Dùibuqǐ, zhètiáo lù yìzhí zài zhèr, nǎr yě bú qù!

Excuse me, where does this road go to?
I'm sorry but this road is always here, it never goes anywhere!

Lǎoshī: Nǐde zuòyè shìbushì nǐ bàba tì nǐ zuòde?
Xiǎo Bǎo: Nǎ yǒu shénme bànfǎ ne?
Māma zǒng shi nàme máng!

Teacher: Did your father do your homework for you?
Xiao Bao: What else could I do? Mum is always so busy!

This is a notice in the window of the local milkbar:

需要 (required)

(一) 女学生，十七岁

(二) 说汉语

(三) 喜欢小孩子

(四) 星期一，星期二，下午四点 - 六点

大卫 Dàwèi
十六岁，
上中学，
说英语，
喜欢小孩子

玛丽 Mǎlì
二十岁，
上大学，
说英语，
喜欢小孩子

Xiǎo Hú is talking to his friend Chén Xiǎotán.

苏 sū
十六岁，
上中学，
说英、汉语，
喜欢小孩子

安娜 Ānnà
十七岁，
上中学，
说汉语，
不喜欢小孩子

187

Allan has been learning Chinese in Australia. He was lucky enough to go to Singapore recently. He decided to keep a list of new Chinese words he came across. He went shopping for what he needed.

Shòuhuòyuán : Nǐ yào mǎi shénme?

Àlún: Nǐmen yǒu méiyǒu běnzi?

Shòuhuòyuán: Yǒu, nǐ yào jǐběn?

Àlún: Sānběn. Duōshǎo qián?

Shòuhuòyuán: Yìběn liùmáo. Yígòng yíkuài bā.

Àlún: Zhè shi liǎngkuài qián.

Shòuhuòyuán: Zhǎo nǐ liǎngmáo. Xièxie.

Àlún: Xièxie. A! Nǐmen màibumài Yīng-Hàn cídiǎn?

Shòuhuòyuán: Duìbuqǐ, wǒmen méiyǒu. Dànshì wǒmen yǒu Yīngwén cídiǎn.

Anne Miller is an Australian doing business in Beijing. On a free afternoon she goes exploring. At one point she has to ask the way — and gets two responses.

APPENDIX 3: ENGLISH TRANSLATION OF TEST PAPER AND MAGAZINE

QUESTION BOOK

Magazine page numbers of the relevant stimulus materials are indicated. Keys are marked with asterisks.

Example

It is now .... o'clock.

A six B three C four D ten

Page 1

1 She has a

A little brother.* B little sister.
C big sister. D big brother.

2 Her mother and father

A are not both teachers. B are not also.
C are both teachers.* C both are not teachers.

Page 1

4 What is on the car?

A There is a person. B There is a cat.*
C There is nothing. D There is a dog.

5 Where is the ball?

A behind the car B wrong
C in front of the car* D under the dog

6 How many people are there behind the car?

A some B a dog
C one* D three

7 Is there a ball on the car?

A There are three. B There is.
C There is a person. D No (there is none).*

Page 2

8 What day is today?

A Thursday* B Wednesday
C four days D 6 o'clock

9 Yesterday was

A 1989. B the seventh (number 7).
C Wednesday.* D 1991.

10 Tomorrow is

A	5th January.	B	Friday.*
C	today.	D	1991.

Page 4

20 Allan wants to

A	buy note-books.*	B	give note-books.
C	read note-books.	D	look for note-books.

21 He gives the shopkeeper

A	60 cents.	B	$2.
C	20 cents.	D	$10.

22 As change, the shop-keeper gives Allan

A	$1.80.	B	60 cents.
C	$1.	D	20 cents.*

23 Allan also wants to buy

A	an English dictionary.	B	a writing brush.
C	an English-Chinese dictionary.*	D	Chinese characters.

MAGAZINE

Page 1

My name is Liu Xiaohong. I live in Shanghai. There are four people in my family: Dad, Mum, my little brother and me. My dad is a teacher; so is my mum. I am fourteen. I go to high school. My little brother is a primary school student. He loves playing football. I don't like it; I like playing volleyball.

Page 2

(Information for farmers - phases of the moon, etc)

6/9/1990

Reminder: 8 o'clock class
Mid-day come home for lunch
4 o'clock football

(Second half of page: no questions attached)

Page 3

Required:

1. 17 year-old female student
2. Able to speak Chinese
3. Likes small children
4. Mondays, Tuesdays, 4pm-6pm

David	Mary
16 years old	20 years old
high-school student	university student
speaks English	speaks English
likes small children.	likes small children

Sue	Anna
16 years old	17 years old
high school student	high school student
speaks English and Chinese	speaks Chinese
likes small children	does not like small children

Boy: What time is it?
Girl: Nearly 11 o'clock.
Boy: It is very hot. Let's go swimming.
Girl: I'm too busy now. Let's go this afternoon, okay?

Page 4

Shopkeeper: What do you want to buy?
Allan: Have you got any note-books?
Shopkeeper: How many do you want?
Allan: Three. How much is that?
Shopkeeper: Sixty cents each. One dollar eighty altogether.
Allan: Here's two dollars.
Shopkeeper: (I'll give you) Twenty cents change. Thanks.
Allan : Thank you. Oh! Do you sell English-Chinese dictionaries?
Shopkeeper: Sorry, we haven't got any, but we have some English dictionaries.

Woman: Excuse me, where is the library?
Man: I don't know.
Woman: Beside the Post Office.

APPENDIX 4: MATRIX OF ITEM CHARACTERISTICS

Item	Logits	WL	CC	OP	Pron	Synt	MM	UW	GFP	Quests	Stim	Locate
1	-1.62	1.33	•	2.33	No	Yes	No	No	Yes	Pinyin	Pinyin	No
2	-.94	1.25	•	1.00	No	Yes	Yes	No	Yes	Pinyin	Pinyin	No
3	-1.95	1.00	•	3.00	No	No	No	No	No	English	Pinyin	No
4	-1.52	1.60	•	3.20	Yes	No	No	No	Yes	Pinyin	•	•
5	0.73	1.40	•	5.40	Yes	No	Yes	No	Yes	Pinyin	•	•
6	0.19	1.60	•	3.00	Yes	No	No	No	Yes	Pinyin	•	•
7	0.08	1.40	•	3.75	Yes	No	No	No	Yes	Pinyin	•	•
8	0.39	2.25	8.66	5.25	Yes	No	Yes	Yes	No	Pinyin	Char	No
9	0.77	2.00	8.33	5.66	Yes	Yes	Yes	Yes	No	Pinyin	Char	Yes
10	0.20	2.00	8.33	5.66	Yes	Yes	Yes	Yes	No	Pinyin	Char	Yes
11	1.19	1.30	7.75	5.75	Yes	Yes	Yes	Yes	Yes	English	Char	Yes
12	-0.22	1.00	10.00	4.00	No	No	No	No	Yes	English	Char	Yes
13	-0.76	1.50	3.30	6.00	No	No	Yes	Yes	No	English	Char	No
14	-0.19	1.63	7.40	3.00	Yes	Yes	No	No	No	English	Char	No
15	0.99	1.40	5.14	3.80	Yes	No	Yes	No	Yes	English	Char	No
16	0.74	1.50	5.33	4.50	Yes	Yes	No	Yes	Yes	English	Char	Yes
17	0.49	1.66	6.60	4.33	No	Yes	No	No	No	English	Char	Yes
18	0.07	1.63	6.60	4.33	Yes	Yes	No	No	No	English	Char	Yes
19	0.95	1.50	7.60	3.00	No	Yes	No	No	No	English	Char	Yes
20	-0.43	1.60	•	4.33	No	No	No	Yes	No	Pinyin	Pinyin	Yes
21	0.75	1.75	•	7.25	Yes	Yes	No	Yes	No	Pinyin	Pinyin	No
22	1.07	1.60	•	7.80	Yes	No	Yes	Yes	No	Pinyin	Pinyin	No
23	0.60	1.16	•	6.00	No	Yes	No	Yes	No	Pinyin	Pinyin	No
24	-0.91	1.33	•	1.33	No	Yes	No	No	No	English	Pinyin	No
25	-0.67	1.63	•	7.00	No	Yes	Yes	Yes	Yes	English	Pinyin	No

Key to Item Characteristics

CC	Mean character complexity (stroke-count)
WL	Word-length (mean number of morphemes per word)
MM	Presence of vocabulary with multiple meanings
GFP	Presence of vocabulary which has been presented in the same unit as other vocabulary from the same semantic field
PRON	Presence of vocabulary with apparent difficulty of pronunciation
SYNT	Presence of syntactic patterns unlike English patterns
OP	Mean unit number of introduction of lexical items
UW	Presence of unknown words
QUESTS	Language of item stem and alternatives
STIM	Script-style of input text
LOCATE	Task-type requiring candidate to locate relevant part of text
CHAR	Character text

TEST INTERPRETATION, TEST USE, AND PEDAGOGICAL IMPLICATIONS

Micheline Chalhoub-Deville
University of Minnesota

ABSTRACT

Test score interpretation and use are the staple of construct validity. As such, in addition to the concern with measurement accuracy, it is imperative that the meaning of test scores and their intended use(s) be also documented. Along these lines, qualitative speech analyses are undertaken in the present paper to help in the interpretation of the dimensions underlying student performance on oral tasks. Results of these analyses yield rich information that explicate the meaning of the dimensions by delineating their specific features as manifested in the speech samples. Also discussed in the paper are the ramifications of these results for pedagogical use. Insights that linguistic accuracy and communicative skills in general, and their specific features specifically, provide for instructional material and activities are addressed. Furthermore, a case is made for curricular improvements to help learners develop well-rounded L2 abilities and to improve their use of the language for real-life communication. Finally, with regard to assessment, it is argued that generic assessment criteria do not reflect the critical features operating in a given context, and assessment practitioners are urged to study their contexts of use and to tailor their criteria according to the particulars of those contexts.

INTRODUCTION

The emergent consensus among measurement experts, as documented in the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985), confirms the central role of validity in measurement research. Researchers such as Anastasi (1986, 1990), Guion (1980), Messick (1975, 1989), Moss (1992), and Shepard (1993) view construct validity as the foundation of the validity concept. In addition, these researchers regard test score interpretation and use as fundamental aspects of the construct validation process. With respect to the practice of construct validation, the National

Council on Measurement in Education (NCME) recently released the Code of Professional Responsibilities in Educational Measurement (1995), the product of a three-year effort, which also promotes appropriate test interpretation and use as a primary responsibility of not only NCME members but of anybody involved in educational assessment. In short, test score interpretation and use is a central aspect of construct validity and is an integral responsibility of measurement professionals.

In line with the general measurement field, second language (L2) testing researchers have placed increasing emphasis on construct validation in theory and practice. For example, in the introduction to their book, *A new decade of language testing research: Collaboration and cooperation*, which includes papers from the 1990 Language Testing Research Colloquium, Douglas and Chapelle (1993) underscore that

> for tests to be used appropriately, test developers must demonstrate, and test users must consider, evidence concerning the meaning of test scores
>
> (Douglas and Chapelle 1993:10).

Hence, L2 measurement professionals must be concerned with the interpretation and use of test scores, in addition to ensuring that these scores are accurate measures of students' proficiency. The present paper focuses on explicating the meaning of dimensions that emerged as underlying students' performance on various oral tasks and considers the implications of the dimensions for pedagogical practices. First, the paper briefly describes previous research that reported the quantitative findings regarding the dimensions salient to native speakers when assessing the performance of L2 learners. The paper proceeds to analyse the speech samples used to derive the dimensions and to illustrate the defining features of these dimensions. Finally, the paper discusses the ramifications of the findings for instruction and for assessment.

PROCEDURES AND ANALYSES

Speech samples

Six students provided the speech samples examined in this study. The six students, two males and four females, ranged in age from 23 to 33.

These students were studying Arabic as a foreign language (AFL) at a land-grant midwestern university in the United States. All six students had completed at least four quarters of modern standard Arabic (MSA), and at the time of the study were enrolled in an intermediate level Arabic language class.

Students were audio-taped in a recording studio while performing three tasks: a modified oral proficiency interview, a narration of a sequence of six pictures depicting a story, and a read-aloud of a short printed text. These tasks were chosen to elicit a wide range of the students' L2 oral ability. These tasks were administered in conjunction with the final test for that class. Students received credit points for taking these tasks.

The purpose of the original research (Chalhoub-Deville, 1995a) was to derive the dimensions underlying scores provided by native speaker rater groups when assessing learners' L2 oral proficiency. The three rater groups, all native speakers of Arabic, included: (a) AFL teachers practising in the U.S.; (b) non-teaching raters who had been residing in the U.S. for at least one year; and (c) non-teaching raters living in Lebanon. The raters evaluated the speech samples of AFL students performing the three tasks.

The individual raters provided two types of ratings: (a) a holistic rating of the quality of each of the speech samples; (b) ratings for each speech sample on specific scales representing linguistic and communicative criteria typically used in L2 oral assessment. The holistic rating data were analysed using INDSCAL, a multidimensional scaling (MDS) model, to derive the underlying dimensions and their salience to each of the three rater groups. Based on the INDSCAL analysis of the holistic scores, three dimensions were derived. After deriving the dimensions, mean ratings of each of the specific scales were regressed on the speech samples' dimensional coordinates to facilitate the interpretation of those derived dimensions. Regression indices indicated that the first dimension was best represented by the two specific scales "grammar" and "pronunciation," the second dimension by "creativity" and "adequacy of information in student's narration," and the third dimension by "giving detail unassisted" and "length of student's responses." The three dimensions were labelled "grammar-pronunciation," "creativity in presenting information," and "amount of detail provided," respectively.

Although the regression analyses were helpful in identifying the specific scales relevant to the interpretation of the derived dimensions, the qualitative analyses were critical not only to corroborate the regression findings but also to yield a more detailed interpretation of these dimensions. The qualitative analyses were essential because they pinpointed the salient attributes in the students' speech samples and, thus, helped explicate the features of the derived dimensions. The following sections present examples of the qualitative speech analyses which were performed in order to assist in the interpretation and the explication of the three derived dimensions. In addition, the significance of the finding from the speech analyses for instruction and assessment purposes are discussed.

Grammar-Pronunciation

The results of the regression analyses point to a single dimension being represented by both "grammar" and "pronunciation". While initially this might seem perplexing, the qualitative analyses make the combination quite clear. The speech sample analyses, and more specifically the types of student error, indicate that the association between grammar and pronunciation results from the inflectional markers in Arabic. The following examples illustrate the point.

Example 1 is from student one's speech sample. The sentence "qarrara an yarkabu darraajatuhu," means "he decided to ride his bicycle."

	English Translation	Arabic Utterance	Student Utterance	Accuracy
1	He decided	qarrara	qarrara	correct
2	to ride	an yarkaba	an yarkabu	incorrect
3	his bicycle	darraajatahu	darraajatuhu	incorrect

(he rides=yarkabu)
(bicycle=darraajatu; his=hu)

As the above analysis indicates, the first phrase in student one's sentence is correct, but the second is incorrect. By saying "an yarkabu" instead of "an yarkaba," the student fails to make the appropriate inflectional change. The inflectional marker "u" at the end of the

student's second utterance "yarkabu," meaning "he rides," indicates that the verb is in the present tense. When a present tense verb, however, is preceded by the particle "an," as in this example, the inflectional marker /u/ should change to /a/. In the third phrase, the student mispronounces "darraajata-hu" as "darraajatu-hu". Usually, when nouns such as "darraajatu" are in nominal position they take the nominal case marker /u/. However, when such nouns are in the accusative, as in the present example, then the ending changes from /u/ to /a/. In this instance the student does not make the appropriate inflectional change to the noun.

There are similar indications of this association between grammar and pronunciation in the speech analysis of student three. The association presented here concerns nouns. The utterance analysed is "shawaarii` muzdaHim walnaas Haar walmadiina jamiila walahraamaat `Ziima," which means "[Cairo has] busy streets, and the people are hot, and the pyramids are great."

	English Translation	Arabic Utterance	Student Utterance	Accuracy
1	busy streets muzdaHima	shawaarii` muzdaHim	shawaarii` incorrect	
2	and the people are hot	walnaas Haara	walnaas Haar	incorrect
3	and the city is beautiful	walmadiina jamiila	walmadiina jamiila	correct
4	and the pyramids are great	walahraamaat `Ziima	walahraamaat `Ziim	incorrect

As the above analysis shows, student three mispronounces phrases 1, 2, and 4, but produces the third utterance correctly. The correct third utterance indicates that the student knows of the gender agreement rule between nouns and adjectives. In the third utterance, the word "city," i.e., "madiina" is feminine and the student supplies the correct feminine inflectional marker /a/ for the adjective. The recurring problem in the phrases 1, 2 and 4 is due to the fact that the student does not provide the correct adjectival form for plural masculine nouns. In Arabic, when masculine nouns are pluralized they are treated grammatically as feminine nouns. Consequently, adjectives accompanying plural nouns, even if those nouns are masculine, should be in the feminine form. In

all three phrases, the student repeatedly fails to produce the required feminine inflectional marker /a/, thus rendering masculine adjectives for plural nouns.

Another example from student four's speech exemplifies how an inflectional marker mispronunciation causes a verb tense mistake. The sentence is "hadatha iSTidaam," meaning "an accident happened."

	English Translation	Arabic Utterance	Student Utterance	Accuracy
1	happened	hadatha	hadathu	incorrect
2	an accident	iSTidaam	iSTidaam ma'	correct

The verb in the above sentence is in the past tense, and past tense verbs typically end with the inflectional marker /a/ and not /u/. By saying "hadatha" instead of "hadathu," student four makes a tense mistake.

The above examples focus on the association between grammar and pronunciation. These examples show that inflectional markers in MSA are such that the smallest mispronunciation is not only an error in pronunciation, but can also be an error in verb tense, gender, etc. thus explaining how grammar and pronunciation function jointly.

Creativity in Presenting Information

Because the two speech attributes, "creativity" and "adequacy of information in student's narration," also do not readily combine to produce an obvious interpretation of dimension two, the qualitative analysis of students' speech samples is again critical. Examples from the speech sample analysis showing the defining aspects of this dimension are provided below. Student one's narration explicates several attributes of dimension two. The following is a translation of her narration:

> It was a beautiful day and there was a young man who has a new bicycle; and he decides to ride his bicycle; and he was riding his bicycle when he saw a beautiful girl; and he watches the girl and does not watch the road; and the poor young man broke his leg and went to the hospital and the beautiful girl came and gave him flowers and the young man said: "thank you very much my love," and she said "welcome," and he said: "Oh, I love you very much." And after that they got married.

To begin, student one describes the pictures using a narrative approach to which the listener can immediately relate. The student begins her story by saying "it was a beautiful day and there was a young man..." She indicates the end of the story by prolonging the /a/ vowel in the final position of the last word, "tazawajaa," i.e., "got married," a characteristic strategy used by a narrator to end a story in Arabic.

Additionally, in an attempt to enliven her account, student one adopts varied discourse styles and language codes to tell the story. Together with the narrative, she incorporates a dialogue style in one segment of her sample. It is also interesting that while the student uses MSA throughout when narrating, she switches to colloquial Egyptian for the dialogue segment. The switch is not only more appropriate for communication (see the discussion on diglossia below), but also more natural and amusing for the native Arabic-speaking listener. Such a switch contributes significantly to the enjoyment of the story.

Student two also provides an engaging account of the pictures and his narration explicates certain defining features of the "creativity in presenting information" dimension. Provided below is the translation of student two's narration:

> In the first picture, the man takes something from his bicycle. Perhaps the thing is a kind of security. The man rides his bicycle on the road beside a lake. The man sees a beautiful girl. She is on her bicycle too; and they smile; and he is still thinking about her; and an accident almost happens between the man and the telephone pole; and he is in bed in the hospital now because the girl is visiting him and she brought him beautiful flowers. The man rides the bicycle once again and with him is the girl; and they are smiling. End of the story.

As the speech sample shows, student two embellishes his account to capture the listener's attention. The stylistic strength of student two's narration is in the crisp descriptions, e.g., "on the road beside a lake," "sees a beautiful girl," "he is still thinking about her," and "they are smiling." Such description about the setting and how the characters think and feel provide the listener with a vivid feel for the story. Furthermore, although student two's story telling tends to reflect less the narrative approach, such elements exist. For example, after describing the couple riding away together smiling, student two declares the end of the story.

Whereas the previous examples focus on the creative features in students' speech samples, the following example from student five's narration portrays the dearth of such creativity. The translation of student five's narration sample follows:

> A month ago, I decided to ride in my bicycle to the road; and I rode in my bicycle in a road; and I saw a very beautiful girl and I did not see the tree; and I was [long pause]. I stayed in a hospital since a week ago; and the girl I saw in the road visited me; and now girl and I rode in our bicycles all the week.

Overall, student five provides quite a confusing and bland account of the story. The student, who is a female, projects herself as the male character in the story and narrates the story from the first person male perspective. Additionally, her account provides a tedious depiction of the story line. The student presents hardly any embellishing elements to enrich the story and enable the listener to project himself/herself into the story and to connect with the characters.

The fact that the above examples focus on the narration samples is not to indicate that "creativity in presenting information" is not meaningful to other tasks, such as the interview. As an example, student four provides a clear illustration of "creativity in presenting information" in the interview. In general, student four's responses to the interviewer's questions are appropriate, colourful, and engaging. For example, after talking about her academic plans and in response to the question of whether she would like to get married one day, she says "God willing, although this is more difficult than studying." As such, the student shows wit and character in answering the question that engage the listener. In short, the speech sample analyses explicate "creativity in presenting information" as focusing on using appropriate and creative discourse.

The Amount of Detail Provided

In a similar way to the previous two dimensions, examining the speech samples qualitatively explicates the meaning of the "amount of detail provided" dimension. The following analyses focus on the performances of students one and three. In the interview, student one provides very short responses. The following is a translated excerpt of student one's interview:

Interviewer: Have you visited a European country?
Student one: Yes, France.
Interviewer: How long did you stay in France?
Student one: Two months.
Interviewer: In what city?
Student one: Toulouse.
Interviewer: Is it pretty?
Student one: Yes, very pretty.
Interviewer: Is it prettier than Paris?
Student one: No, Paris is prettier.

As evidenced, the student does not elaborate or volunteer any additional information. Thus, the interviewer is forced to interject repeatedly to elicit details.

Student three, on the other hand, repeatedly explicates his responses by elaborating and including details. Below is a translated segment of student three's interview:

Interviewer: Did you like the city of Cairo?
Student three: Yes, very much! Busy streets, and the people are hot and the city is beautiful, and the pyramids are great...
Interviewer: Do you work while studying at the University?
Student three: Yes, I work in an average restaurant called A La Carte. I work in the kitchen two days, and one day I work in the dining hall – server. I am a server one day. But during the other days of the week I work at a centre for adolescence, Huckleberry – Huckleberry House...

As this segment shows, student three tends to respond in detail to the interviewer's questions. In responding to a question about whether he liked Cairo, he does not simply answer in the affirmative. He proceeds to explain the reasons. Similarly, when asked about his work, this student names the places where he works, provides his work schedule and explains his duties. In contrast to student one, the interviewer here is not forced to interject in order to encourage the student to provide detail to his responses. The student elaborates his responses unassisted.

Other examples illustrating differences in attention to detail focus on the narration samples of students one and three. The pattern of detail in student one's and three's narrations is similar to that observed in their interview samples. In narrating, student one presents a more summarised account of the visual prompt, whereas student three describes every

picture at length. In describing the accident, for example, student one says "and he was riding his bicycle when he saw a beautiful girl. And he watches the girl and does not watch the road..." (translation). In narrating the same segment, on the other hand, student three elaborates in his description. Provided is a translation of the segment:

> In the next picture, he rides his bicycle beside the lake on not a good road; and then he looks at a beautiful girl wearing in short pants; and then after that an accident happened with a big tree—telephone trees. He rides outside the road.

In short, student three's performance, on both the interview and the narration, reflects attention to detail. Student one's speech, on the contrary, is brief and contributes minimal detail. In conclusion, similar to other dimensions, the speech sample analyses explain the meaning of the "amount of detail" dimension by delineating specific features.

PEDAGOGICAL USE

The present speech analyses have important pedagogical consequences. The analyses contribute significant insights into instructional and assessment practices. These insights are discussed under the headings of linguistic accuracy, communicative skills, and assessment.

Linguistic Accuracy

The speech sample analyses ascertain the central role "grammar-pronunciation," (ie. inflectional markers) play in linguistic accuracy in Arabic. Clearly this translates into the necessity for depicting such grammatical features in instructional materials and activities. In presenting these features, this raises the issue of which is the appropriate approach to adopt. Before addressing this, two questions need to be discussed. First, is "grammar-pronunciation" particular to Arabic? Second, what role does linguistic accuracy play in the development of L2 ability?

With regard to "grammar-pronunciation," it is important to note that such a linguistic feature is not restricted to Arabic. Italian and Spanish, for example, share such a feature with Arabic. "Este" in Spanish depicts how grammar and pronunciation are related. With the

demonstrative adjective "este," as in "este libro," i.e., "this book," the natural stress pattern results in the first syllable receiving primary emphasis. When the natural stress is on the second syllable, however, "esté" (in the subjunctive mood) means "he, she, or it were." For example, "Espero que el hombre esté aqui," means "I wish that the man were here." Thus, depending on the word stress placement, the function of the word changes.

Another example illustrating the relationship between pronunciation and grammar focuses on verbs. For all regular "ar" verbs in Spanish, a change in vowel stress produces a change in person and in verb tense. For example, "estudio" means "I study." When the final syllable receives primary stress, i.e., "estudió," the verb means "he," "she," or "it studied." As such, a change in vowel stress causes the person to change from first to third person and the verb tense from present to past (preterit). In conclusion, the "grammar-pronunciation" features are not limited to Arabic and thus the following discussion is also relevant to other languages.

The second issue which needs addressing is the role played by linguistic accuracy in the development of learners' L2 ability. While many L2 educators have been promoting a more communicative instructional approach, a number maintain that training students in linguistic accuracy is paramount. This as yet unreconciled tension between the linguistic and the more communicative aspects of the language has often led to discrepancies in instructional materials and practices. For example, Taha (1995), who discusses this discrepancy issue in the AFL context, points out that the communicative content of books is often redefined to serve as extensive grammar explications. She also points to a mismatch between the presentation of grammatical structures in books and the frequency of those structures in real-life communicative language use, and documents a lack of a coherent presentation of grammar to students at various levels of study. Such a disjointed treatment of grammar in materials and in practice can only confuse students.

The issue is not whether curricula and teaching need to emphasise grammatical or communicative skills. The present analyses and arguments underscore that both are central to communication. Grammar

features should be presented throughout as necessary means for the construction of meaning. As Kramsch (1991) argues, the issue:

> is not whether metalanguage is necessary or not, but how to introduce metadiscursive consciousness at all levels of the curriculum
>
> (Kramsch 1991:202).

Educators, therefore, need to focus on exploring approaches to teaching linguistic accuracy that consider how students internalise grammatical structures, including inflectional markers, for the construction of meaning.

Traditionally, in developing students' linguistic accuracy, educators have relied on production-based activities, e.g. drills. Upon reviewing the literature on language acquisition, Ellis (1995) argues for an approach that incorporates activities where the focus is on the L2 input to which the student is exposed. He provides the following rationale for such type of activities:

> Interlanguage development can be more readily influenced by manipulating input than output, an approach that requires interpretation tasks that cause learners to attend to specific grammatical properties in the input, to identify and understand the meanings they convey, and to compare the form-function mappings of the target language with those that characterise the interim stages of learners' own interlanguage development.
>
> (Ellis 1995:99)

In other words, unlike production-based activities where students are encouraged to automatize grammatical features without analysis, interpretation tasks require learners to deliberately process grammatical forms and to understand how these forms serve communicative functions. Rutherford (1987) advocates a similar approach using consciousness-raising types of activities.

Such an input based approach is appropriate for the teaching of inflectional markers. For example, both students one and three seem to have problems with markers for verbs, and instruction can incorporate, in addition to the prevalent practice of drilling to automatise the conjugation of the verb, the interpretation type of activities to make these students attend to these problems. The teacher can draw students' attention to these mistakes directly through error correction. The purpose

of such feedback is to encourage these students to notice and develop awareness of the appropriate tense markers and their variations when preceded by particles. The teacher can also design activities that assist learners to monitor their use of these verbs. In conclusion, in recognising the central role of grammar for constructing meaning, the focus should shift from *whether* curricula need to encompass this linguistic aspect to *how* to teach it by manipulating both the output and the input.

Communicative Skills

The present findings point to learners' need to develop their linguistic as well as their creative and functional skills, aided by appropriate materials and instructional practices. While the Arabic field has made important strides to facilitate the development of learners' communicative skills, much of the Arabic instructional material specifically, and instructional material in general, still consists of heavily structured activities that exert excessive control over learners' communicative production. Many so-called communicative activities incorporate display questions by the teacher to which students, one at a time, are expected to supply answers. Another popular activity requires students to practice some prestructured exchange in pairs or small groups where students typically focus on what the teacher wants them to say. Such activities are heavily controlled and do not allow students to create their individual expression to realise their intended message. As Kramsch (1991) points out, the communicative language approach is "predicated on a unique, original, creative individual exchanging and clarifying meanings he or she clearly intended" (p. 195). In short, it is important to provide students with real-life simulations where they can practice their communicative skills unmediated by the teacher's own expression.

Teachers need to help students improve the way they express themselves both linguistically and stylistically to achieve control over various speech functions within and across discourse styles. An L2 program needs to include a range of activities such as narration, debate, and conversation where students practice various manipulative, heuristic, ideational, and imaginative functions (Bachman, 1990). Elgibali and Taha (1995) describe strategies intended to help learners improve their functional skills as per the conventions governing different discourse types such as the ability to carry on a conversation:

a. When asked a question, answer it and then ask another one, e.g., "How are you?" "Fine, thanks. And you?"

b. When asked a question, answer it giving extra information, e.g., "are you a student?" "Yes, I'm studying history. How about you?

(Elgibali and Taha 1995:91)

Student one who is described in "The Amount of Detail" section as being very brief in her interview responses could benefit from a similar kind of activity where she could practice elaboration and turn-taking, useful skills in conversation. In conclusion, creating opportunities where students are challenged to achieve and improve their use of the language for real-life communication is pedagogically sound and should be the mainstay of the L2 classroom.

The present discussion of the communicative skill aspects would be incomplete without addressing the issue of diglossia brought up when describing the narration sample of student one. To recap, student one used MSA when narrating, and colloquial Arabic when she created the dialogue between the young man and the girl.

The debate as to whether to teach MSA or colloquial is spirited. In teaching AFL, educators have typically employed MSA, and colloquial Arabic has remained largely ignored. The conventional argument is that students are better served if the classroom focuses on MSA, because that will enable them to communicate with Arabic speaking people from various countries and regions. Proponents of this approach usually advocate either postponing the teaching of colloquial until later, e.g. third year of studying Arabic, or relegating the learning of colloquial to outside the AFL classroom, for example, to be acquired in the target community. This almost exclusive focus on teaching MSA, however, violates the sociolinguistic aspect of the language as it is used in the Arabic-speaking communities. Moreover, such an approach does not prepare students to function appropriately in an Arabic-speaking context.

The position of integrating both MSA and colloquial into the curricula is supported by arguments made by several educators (Al-Batal, 1995; Belnap, 1987, 1995; Elgibali and Taha, 1995; and Younes, 1995). Elgibali and Taha (1995), for example, argue that because learners will be using a sociolinguistically inappropriate language code and

> any attempt to use *fusha* [standard Arabic] in mundane situations (like shopping) would immediately provoke social ridicule.
>
> (Elgibali and Taha 1995:97)

Furthermore, supported by the results of his survey of 568 AFL students, Belnap (1995) contends that suspending the teaching of colloquial disregards the students' stated needs to develop their oral skills to communicate with speakers of Arabic. He also points out that educators cannot afford to wait until the third year to teach colloquial, since the majority of the students, 80% of those surveyed, will have dropped out by that time. In short, a coherent instructional approach to teaching AFL requires matching the Arabic varieties with their context of use as observed in real-life situations. These arguments make a strong case for AFL curricula to incorporate such an integrated system of communication from year one. For a review of approaches for incorporating MSA and colloquial into the curriculum, see Al-Batal (1992, 1995), Alosh (1991), Ryding (1995), and Younes (1995).

The present analyses and the discussion in this section explicate various communicative aspects which are necessary for students to develop in order for their language use to resemble that encountered in real-life. Consequently, L2 material and activities should incorporate such communicative aspects so that learners have a greater chance of functioning successfully outside the classroom.

Assessment

Assessment is an integral aspect of the pedagogical process. The recommendations made above, which emphasise both linguistic accuracy and communicative skills in classroom practice, translate into a call for a comparable emphasis of these components in assessment as well. Moreover, the quantitative and qualitative analyses underscore certain aspects of these linguistic accuracy and communicative components, i.e. "grammar-pronunciation," "creativity in presenting information," and "amount of detail provided," as salient for assessment in the present context.

The context, with its particular purpose, language, examinee, task, test score user, etc., renders certain aspects of the construct critical and others irrelevant in a particular context (Chalhoub-Deville, forthcoming). Whereas the L2 testing literature has a long history of documenting the

various issues pertaining to some of these variables, (for example the purpose of assessment), the discussion of others, such as, the end-user of the results, is scarce. With test use as a fundamental aspect of validity, addressing the issues pertaining to the end-user is pressing. The following discussion explores the end-user when assessment is intended for classroom use and when it is performed for use outside the classroom. (This distinction was suggested by Wesche, 1992.)

In terms of assessment for classroom use, where students are being tested in order to inform further instruction, teachers are typically the end-users of the test results. Assessments, therefore, should be congruent with the criteria deemed important by the teachers. For example, when the test is intended to measure learners' ability to purseue studies where the L2 is the medium of instruction, tests should incorporate the concerned teachers' perceptions of learners' ability to function in that language. Alderson and Clapham (1995) and Upshur and Turner (1995) similarly argue that in order to obtain valid information of students' performance, classroom assessment should include teachers' views and beliefs. In other words, if teachers are not involved in the development of assessment for instructional purposes, test score interpretation and use is jeopardised because there are no guarantees that scores obtained will reflect teachers' perceptions.

In terms of assessment for use outside the classroom, where students are being tested, for example, to evaluate their ability to communicate with native speakers or for professional certification, test scores should take account of the views held by representatives of that population and not those of the L2 teachers. This recommendation is based on both theoretical arguments and research findings (Barnwell, 1989; Brown, 1995; Chalhoub-Deville 1995a, 1995b; Galloway, 1980; Hadden, 1991; Ludwig, 1982) indicating a disparity between the two groups in terms of their perception of L2 proficiency. Ludwig (1982), for example, contends that because teachers are continually exposed to learners' interlanguage, their proficiency perceptions are not consistent with those held by non-teachers. Likewise, Barnwell (1989) argues that teachers tend to match their assessment of L2 learners' proficiency to the models with which they have been trained. Brown (1995) advances a similar argument when investigating professional certification of tour guide candidates. Upon observing the rating behaviour of a group of teachers and a group of tour guide professionals, Brown observes that,

in the given context, the perceptions of the tour guide representatives are likely to be more appropriate than those of the L2 teachers. In short, the literature documents that teachers and non-teachers vary in their professional training and background experiences and, thus, differ in their expectations and evaluations of students' proficiency. Because ratings reflect raters' own assessment criteria, the disparity between the two groups indicates that a mismatch between the intended use of assessment and the rater group employed threatens the validity of test score interpretation and use.

With regard to this study, the salience of the three dimensions to the rater groups described in the Procedures and Analyses section, previous research (Chalhoub-Deville, 1995a) shows that "creativity in presenting information" is salient to the teachers, "amount of detail provided" is emphasized by the non-teachers' residing in the U.S., and "grammar-pronunciation" is most relevant to the non-teachers living in Lebanon. A holistic score based on teachers' ratings, therefore, is communicating different information from that provided by either the non-teachers in the U.S. or the non-teachers in Lebanon. This discrepancy in dimensional emphasis among the various rater groups reinforces the significance of closely matching the rating scheme with the intended use of assessment.

In conclusion, the present discussion underscores the importance of differentiating between assessment employed to evaluate students' proficiency for classroom use and that for use outside the classroom. To help ensure appropriate interpretation of results, test developers and researchers are encouraged to include the perceptions and criteria espoused by representatives of the group(s) most consistent with the context(s) within which those assessment scores are used.

CONCLUSION

Test score interpretation and use are the staple of construct validity. In focusing on these critical aspects of construct validity, qualitative analyses are undertaken in the present paper to corroborate and explicate the meaning of the quantitatively derived dimensions underlying student performance on oral tasks. Such an incorporation of both quantitative and qualitative analyses results in a more rigorous research approach that affords a richer documentation of the meaning of test scores, and contributes to a more informed use of those scores.

Analyses verifying the meaning of test scores are critical for pedagogy. Proper interpretation of test scores influences the development of appropriate instructional materials and activities to facilitate students' development of language ability. Furthermore, with regard to assessment, it is argued that generic assessment criteria jeopardize the validity of scores. Accurate interpretation and use of test scores necessitates the inclusion of criteria that correspond to the perceptions of the end-users.

REFERENCES

Al-Batal, M. (1995) Issues in the teaching of the productive skills in Arabic. In M. Al-Batal (ed.) *The Teaching of Arabic as a Foreign Language* Provo, UT, American Association of Teachers of Arabic.

Alosh, M. (1991) Arabic diglossia and its impact on teaching Arabic as a foreign language. In G. Ervin (ed.) *International Perspectives on Foreign Language Teaching*. Chicago, IL, National Textbook Company.

Alderson, C. & Clapham, C. (1995) Assessing student performance in the ESL classroom. *TESOL Quarterly*, 29, 184-187.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985) *Standards for Educational and Psychological Testing*. Washington, DC, Author.

Anastasi, A. (1986) Evolving concepts of test validation. *Annual Review of Psychology*, 37: 1-15.

Anastasi, A. (1990) Ability testing in the 1980's and beyond: Some major trends. *Public Personnel Management*, 18: 471-485.

Bachman, L. (1990) *Fundamental Considerations in Language Testing*. Oxford, Oxford University Press.

Barnwell, D. (1989) 'Naive' native speakers and judgements of oral proficiency in Spanish. *Language Testing*, 6:152-163.

Belnap, K. (1995) The institutional setting of Arabic language teaching: A survey of program coordinators and teachers of Arabic in U.S. institutions of higher learning. In M. Al-Batal (ed.) *The Teaching of Arabic as a Foreign Language*. Provo, UT, American Association of Teachers of Arabic.

Brown, A. (1995) The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, 12:1- 15.

Chalhoub-Deville, M. (forthcoming) Theoretical models, assessment frameworks, and test construction. *Language Testing*.

Chalhoub-Deville, M. (1995a) Deriving oral assessment scales across different tests and rater groups. *Language Testing*, 12:16-33.

Chalhoub-Deville, M. (1995b) A contextualized approach to describing oral language proficiency. *Language Learning*, 45: 251-281.

Douglas, D. & Chapelle, C. (1993) In D. Douglas & C. Chapelle (Eds.), *A New Decade of Language Testing Research*. Alexandria VA, Teachers of English to Speakers of Other Languages Inc.

Elgibali, A. & Taha, Z. (1995) Teaching Arabic as a foreign language: Challenges of the nineties. In M. Al-Batal (Ed.), *The Teaching of Arabic as a Foreign Language*. Provo UT, American Association of Teachers of Arabic.

Ellis, R. (1995) Interpretation tasks for grammar teaching. *TESOL Quarterly*, 29: 87-105.

Galloway, V. (1980) Perceptions of the communicative efforts of American students of Spanish. *Modern Language Journal*, 64: 428-433.

Guion, R. (1980) On trinitarian doctrines of validity. *Professional Psychology*, 11:385-398.

Hadden, B. (1991) Teacher and non-teacher perceptions of second-language communication. *Language Learning*, 41, 1-24.

Kramsch, C. (1991) The order of discourse in language teaching. In B. Freed (ed.) *Foreign Language Acquisition Research and the Classroom*. Lexington MA, D. C. Heath and Company.

Ludwig, J. (1982) Native-speaker judgments of second language learners' efforts at communication: A review. *Modern Language Journal*, 66: 274- 283.

Messick, S. (1975) The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30, 955-966.

Messick, S. (1989) Validity. In R. L. Linn (ed.) *Educational Measurement* (3rd ed). New York, American Council on Education/Macmillan.

Moss, P.A. (1992) Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, 62:229-58

National Council on Measurement in Education. (1995) *Code of Professional Responsibilities in Educational Measurement*. Washington, DC: Author.

Rutherford, W. (1987) *Second Language Grammar: Learning and Teaching*. NY: Longman.

Ryding, K. (1995) Discourse competence in TAFL: Skill levels and choice of language variety in the Arabic classroom. In M. Al-Batal (ed.) *The Teaching of Arabic as a Foreign Language*. Provo, UT, American Association of Teachers of Arabic.

Shepard, L. A. (1993) Evaluating test validity. In L. Darling-Hammond (ed.) *Review of research in education*, 19. Washington, DC, American Educational Research Association.

Taha, Z. (1995) The grammar controversy: What to teach and why. In M. Al-Batal (ed.) *The Teaching of Arabic as a Foreign Language*. Provo, UT, American Association of Teachers of Arabic.

Younes, M. (1995) An integrated curriculum for elementary Arabic. In M. Al-Batal (ed.) *The Teaching of Arabic as a Foreign Language*. Provo, UT, American Association of Teachers of Arabic.

Upshur, J. & Turner, C. (March, 1995) Task, judge and scale effects in the rating of speaking ability of primary school ESL learners. Paper presented at the Annual Meeting of Language Testing Research Colloquium, Long Beach, CA.

Wesche, M. (1992) Performance testing for work-related second language assessment. In E. Shohamy & R. Walton (eds) *Language Assessment for Feedback: Testing and Other Strategies*. Dubuque, IA: Kendall/Hunt Publishing Company.

INVESTIGATING WASHBACK IN JAPANESE EFL CLASSROOMS: PROBLEMS OF METHODOLOGY

Yoshinori Watanabe
International Christian University

ABSTRACT

In Japan it has long been considered that university entrance exams have a negative impact on teaching and learning of English in schools. Recent research, however, suggests that the relationship between testing and teaching is not so simple and requires fine-grained ethnographic research in order to fully understand its complexities. This paper argues the case for the necessity of conducting empirical research before any conclusions about the presence or absence of washback can be drawn. An example of this type of research is offered and the processes involved in investigating washback in the Japanese context are described in such a way as to provide a model for future research in this area.

INTRODUCTION

The impact of exams, which has been known as 'washback' effects in the ESL/EFL literature, is a widely asserted but rarely attested phenomenon. Watanabe (in progress) has come up with more than 500 general claims concerning the impact of the Japanese university entrance exams on classroom practice, whilst finding only ten empirically conducted studies.

Although these studies provide directions which are helpful in formulating research questions, they do not provide the type of information necessary to innovate education through testing. First, in virtually all the studies surveyed, the notion of 'washback' was not fully conceptualised, nor was its scope clearly defined; therefore, it is often unclear what implications could be drawn for improving education through testing. For example, one of the findings reported by Ariyoshi & Senba (1983) was that third grade high school students took practice tests more frequently than second graders. This finding was used to support their argument that entrance exams were influencing high school education *negatively*, in that the lessons were geared increasingly towards

the coming exams as the grade became higher. However, it cannot be taken for granted that the taking of practice tests is intrinsically harmful. The authors needed to specify more clearly what they meant by 'negative', and to define the role of practice tests in their notion of washback. And before claiming negative effects they would need to examine how the practice tests were used by students and teachers.

Second, since most researchers did not fully analyse the exams which they considered to have an impact, it is not clear whether the observed phenomenon had really been caused by the exam. Fujita (1992) reported that students who were preparing for the university entrance exam tended to learn English by paying particular attention to complex grammatical structures. He speculated that this was because most of the questions employed in Japanese university entrance exams consist of those requiring of testees the ability to understand complex grammar. However, since no in-depth analysis of the exams was reported, his conclusion remains merely a surmise.

Thirdly, most of the studies have university students as subjects. Since these students have already taken the entrance exams, it is not always clear that what the researchers observe is attributable to the effect of the exams. Berwick and Ross, for example, demonstrated that most Japanese students lack motivation to learn English after entering university, and concluded that:

> the intensity of motivation to learn English hits a peak in the last year of high school... Once the university examinations are over, there is very little to sustain this kind of motivation, so the student appears in freshmen classrooms as a kind of timid, exam-worn survivor with no apparent academic purpose at the university
>
> (Berwick and Ross1989: 206)

However, because they dealt with university students, it is not clear from their research whether their students had in fact been motivated during their last year of high school.

Finally, most of the research collected data by indirect means, such as questionnaires and interviews. For example Saito et al. (1984) based their research into the effect of multiple-choice questions on medical education on the self-reports of teachers, and concluded that this type of question distorted teaching in various ways. For example teachers reported that "they tended to teach with an emphasis on the areas that would most likely be covered in the exam" (p. 38 [translation mine]) and

that "students tended to study about details rather than main points," (p.43 [translation mine]). However, because their findings were based on questionnaire data rather than on direct observation, we are left in some doubt as to whether these findings accurately reflect either what was happening in the classroom or how students were preparing themselves for the exams.

In summary, then, the research to date has not been successful for the following reasons; first, the notion of the "effect" of the exam has not been clearly delineated; second, the target exams have not been fully analysed; third, data has been gathered from students after, rather than prior to the examinations; and fourth, conclusions have been drawn on the basis of indirect data collection methods, such as questionnaires and interviews.

While much of the investigation of washback is limited by lack of direct empirical support, there are exceptions. Wall and Alderson (1993), and Alderson and Hamp-Lyons (1994) used classroom-based studies to indicate that washback is not so straightforward a phenomenon as it is generally assumed to be. Wall and Alderson (1993) examined the impact of the Sri Lankan O-level English exam on teaching by observing classrooms before and after the implementation of a new exam and comparing the two sets of data. Results provided evidence for the presence of washback (both negative and positive) on lesson content but not on the method of teaching. The methodology used by teachers after the introduction of the new exam was essentially the same as that observed in the baseline study. They concluded that the exam is only one of the factors that "affect how innovations succeed or fail" (1993: 68). Alderson and Hamp-Lyons (1993), investigating the the washback effect of TOEFL, compared a total of eight non-TOEFL classes to eight TOEFL preparation classes. Results suggested a clear impact of the test on the "curriculum, text books, teachers, attitudes and the content of classes aimed at TOEFL". However, the impact on teaching methodology was much more complex, and in this area they concluded that "the differences between teachers are at least as important as the differences between TOEFL and non-TOEFL" (1993: handout).

To deal with this kind of complexity there is a clear need for different types of observational research in various contexts. The contexts in which tests are used differ greatly, and the type of exam and its potential influence also varies. Test purpose is also variable: some tests

are used to select students; others to place students in classes according to their proficiency levels. Thus, the result obtained in one research area may not be readily generalisable to another context. For example, in Japan, more than 1,000 different university entrance exams are carried out each year, so the results of research relating to a single examination like TOEFL, may not be applicable to the overall situation.

Washback is an extremely complex phenomenon. It can be conceptualised on different dimensions, involving a number of variables, such as teaching, learning, interaction between teachers and students, students' motivation, etc. and therefore needs to be examined from multiple perspectives. Classical experimental research is not appropriate because of the difficulty of controlling for all these variables.

Ethnography, on the other hand may be a more suitable approach to understanding the complex nature of washback. For LeCompte and Preissle (1993: 3) ethnography is characterised by four research strategies. First, ethnography elicits phenomenological data that "represent the world view of the participants being investigated," and "participants' constructs are used to structure the research." Second, in this method, participant and nonparticipant observation is used "to acquire firsthand, sensory accounts of phenomena as they occur in real world settings." Third, ethnographers "seek to construct descriptions of total phenomena within their various contexts and to generate from these descriptions the complex interrelationship of causes and consequences that affect human behavior toward and belief about the phenomena." And fourth, the ethnographic researchers use "a variety of research techniques to amass their data."

The present paper is part of a larger study on the washback effects of the Japanese university entrance examinations (Watanabe, in progress). To demonstrate the value of the ethnographic approach it will focus largely on the research process. Provisional conclusions will nevertheless be presented at the end of the paper.

The paper is in several sections. The first section describes the Japanese educational context which is the setting in which the research was conducted. In the next section, the process of the research is described in several stages, the research questions are identified and the approaches employed to address them discussed. The final section summarises the provisional findings, and identifies the need for further empirical research using methods derived from ethnography.

UNIVERSITY ENTRANCE IN JAPAN

The Japanese university entrance examination system

In Japan the school year runs from April 1 to the end of March. The entrance examinations of most universities are administered in January and February. Each university department produces its own examination which is offered on the campus. As no information concerning the exam (such as scoring methods, weighting of each section, etc., not to mention the intended effect of the exam on teaching and learning) is made public, the most important sources of information for the examinees are the previous year's examinations.

There are three major types of university in Japan; i.e., national, local public, and private. According to the Ministry of Education (1994), out of a total of 534 four-year course universities, 98 (18.4%) were national, 46 (8.6%), local public, and 390 (73%), private. Entry to universities is highly competitive, since more than half of upper secondary level graduates hope to enter tertiary institutions. In 1993, for example, 59.6% of the total of 1,760,000 graduates took the exam; 67% were accepted into university. Thus, as elsewhere, demand outstrips supply. Further, most students who fail to enter their target university undertake a further year of study and take the entrance exam the following year. These so-called *ronin* students contribute to the increase in the number of applicants in the following year. In addition, universities are ranked and there is a widespread belief that entering higher rank universities guarantees better jobs after graduation. Thus university entry is highly competitive.

The teaching of English in Japan

With only a few exceptions (e.g., the natural science-related departments of some universities), virtually all Japanese universities require English or other foreign languages in their screening process.

Secondary schools are expected to follow the guidelines issued by the Japanese Ministry of Education. Various textbooks based on the guidelines are published by private publishers commissioned by the Ministry.

The role of English is gradually changing in Japan; English used to be merely a medium of gaining 'high culture' through the written

language, but now, in reponse to the need for international communication, there is a greater emphasis on active language use involving exchange of information in both speech and writing. To meet these external demands, a new set of guidelines was officially announced in 1989, and implemented in 1994. The first entrance examinations based on the new curriculum will be held in 1997. The new guidelines are more specific than the preceding ones about the nature and range of oral communication skills (including daily conversation, note-taking, giving speeches, and debate), as well as reading and writing to be taught. Given this greater emphasis on communication in EFL teaching, the EFL examinations have been criticised for their lack of communicative content. The need for research into the impact of the exams is called for in such a context.

One of the purposes of this research then was to gather base-line data to allow eventual comparisons to be drawn between the current situation in Japanese EFL classrooms and that which will occur in 1997, once the new curriculum has been introduced.

THE RESEARCH PROCESS

The ideas and terms for the research process adopted here are drawn from Evertson and Green (1986). Figure 1 illustrates the process and the description which follows is based on the diagram.

Received wisdom

My initial interest in washback developed from my experience as a teacher, researcher, and student who had gone through the process of preparing for the university entrance examinations. During this period I subscribed to the prevailing view held in the Japanese society that exams had a negative impact on learning and teaching. It was believed that if the exam were to be improved, then innovation in teaching and learning would automatically follow.

Start questioning

This assumption was called into question when I conducted research into the relationship between language learning strategies and the screening

method of a junior college (Watanabe, 1990). In this study, two types of college students were compared with respect to the range and types of language learning strategies they used. The first type of student had entered the college by taking the entrance examination; the second, by recommendation. The second type of student was completely exempted from the exam, and was accepted on the basis of high school records. It was predicted that the examination students would report using a narrower range of strategies, such as rote-memorisation of vocabulary items, translation, which would benefit their test-taking. The 'recommended' students, on the other hand, were expected to report using a wider range of more communicatively-oriented strategies (eg., learning vocabulary items by using them in actual contexts). The favourable high school records of these students were taken as an indication that they had done well in classrooms where the communicatively-oriented Ministry of Education guidelines had supposedly been implemented. Contrary to these predictions, the result indicated that the entrance examination seemed to have a positive effect on strategy use, in that the exam students reported employing a greater range of strategies than those accepted on the basis of school performance alone, and the type of strategies included communicatively-oriented strategies, such as learning English by using it in actual communicative situations. However, since the research was based on student self-reports rather than direct observation, the reliability of these findings was uncertain. Moreover, it was not clear what had actually been taught in their high school classes. Thus, this was the stage where I felt it necessary to conduct empirical research into washback.

Personal theory

Before designing the research, the initial theorising of washback was first expressed in the form of research questions. The questions raised at this stage were as follows;

- Does washback exist?
- What evidence enables us to say washback exists or does not exist?
- If washback exists, what is its nature (i.e., positive or negative)?
- If washback does not exist, why not?
- If washback exists, under what conditions?

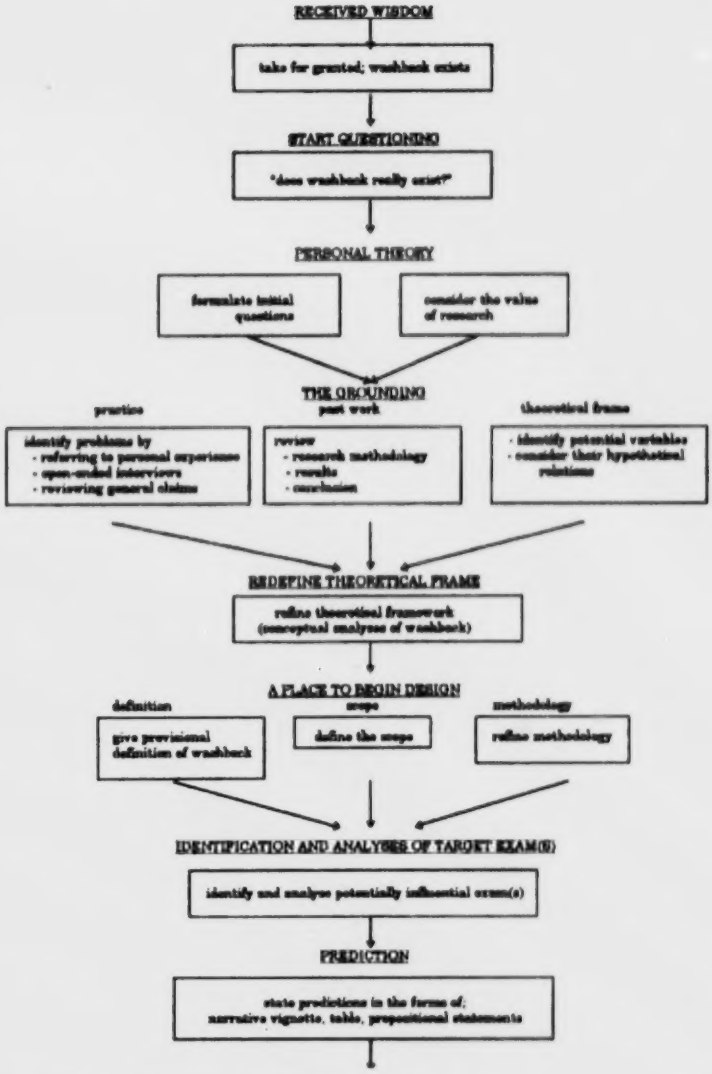


Figure 1: Process of washback research

BEST COPY AVAILABLE

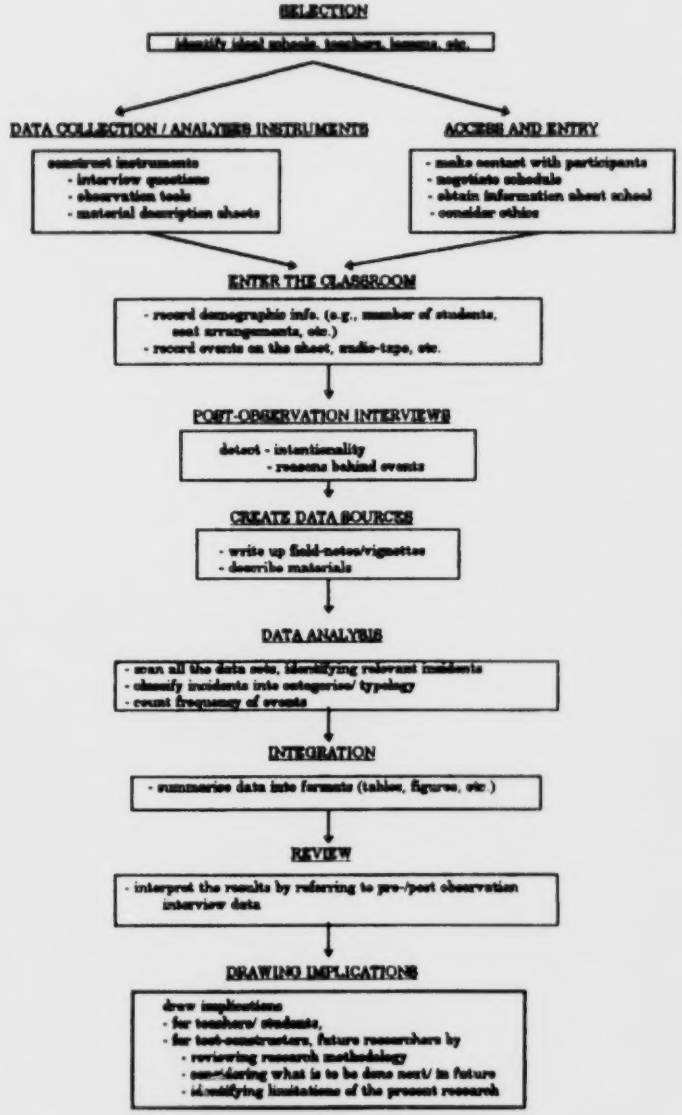

SELECTION
identify ideal schools, teachers, lessons, etc.
DATA COLLECTION / ANALYSIS INSTRUMENTS
construct instruments
- interview questions
- observation tools
- material description sheets
ACCESS AND ENTRY
- make contact with participants
- negotiate schedule
- obtain information about school
- consider ethics
ENTER THE CLASSROOM
- record demographic info. (e.g., number of students, seat arrangements, etc.)
- record events on the sheet, audio-tape, etc.
POST-OBSERVATION INTERVIEWS
detect - intentionality
- reasons behind events
CREATE DATA SOURCES
- write up field-notes/vignettes
- describe materials
DATA ANALYSIS
- scan all the data sets, identifying relevant incidents
- classify incidents into categories/ typology
- count frequency of events
INTEGRATION
- summarise data into formats (tables, figures, etc.)
REVIEW
- interpret the results by referring to pre-/post observation interview data
DRAWING IMPLICATIONS
draw implications
- for teachers/ students,
- for test-constructors, future researchers by
- reviewing research methodology
- considering what is to be done next/ in future
- identifying limitations of the present research

Figure 1: Process of washback research (continued)

BEST COPY AVAILABLE

These five core questions remained unchanged throughout the research.

This stage involved consideration of the value of the research. Clearly research addressing these questions has important implications for Japanese society, since much energy, time and money are spent on the exam at individual, school, and national levels every year. Empirically-based research was seen as an essential investment.

The grounding

Evertson and Green (1986) divide this stage into three components; practice, past work, and theoretical frame.

Practice

The purpose of the practice component was to identify the existing problems with the exams. This information was derived from two main sources; firstly open-ended interviews with the people who had been involved in the entrance examination (including university students, high school teachers, those who were working at special preparatory school called *yobiko*), and secondly, investigation of a range of claims published in magazines, newspaper articles, TV broadcasts, etc. The claims were summarised on note-cards. Each item was then labelled with a tentative title. The source and page numbers were also referred to, and my comments were recorded in the margin. A sample card is shown below:

> [title] negative effects of objective type questions/ positive effects of subjective type questions
>
> ... Our university has recently decided to reform our entrance exam by replacing objective type questions with subjective type, because the objective test is likely to have extremely negative effects on students... (Nakamura, 1985: 11) [translation mine].
>
> [comments] What is referred to by 'subjective' 'objective'; unclear. No evidence is cited.

The note-cards were classified according to the titles, and the most frequently claimed problems were then summarised as sets of assumptions as follows:

Assumption 1: Aural/oral aspects of English are neglected in exam preparatory lessons.

Assumption 2: In exam preparatory lessons, a greater emphasis is placed on de-contextualised mechanical drills rather than communication exercises.

Assumption 3: In exam preparatory lessons, test-taking techniques are taught.

Assumption 4: Exam preparatory lessons are characterised by teaching based on grammar-translation, with detailed explanations about minute grammar points.

Assumption 5: The exam drives students to work harder.

Assumption 6: Different testing methods, such as objective/subjective, and direct/indirect, induce differential washback.

Assumption 7: If the target examination includes listening, then this skill is taught in the course.

Assumption 8: If the target examination requires translation, then the teaching is more likely to be based on grammar-translation.

These assumptions served as "base-line" assertions, around which the subsequent part of the research was organised.

Past work

In the component called past work, the following set of questions was addressed:

- has anything been done in the area?

If yes;

what has been done?

- what methodology has been used?
- what results/conclusions have been drawn?
- what has been lacking in past research?

- would there be any problems particular in the present context (i.e., the Japanese university entrance examinations)?

In addition to the problems identified through the literature review, summarised in the introduction of the paper, there was a problem particular to the context of Japanese educational system. In Japan, as already noted, there are more than 1,000 different types of exams administered every year. Therefore, it was anticipated that identification and selection of target exams for the research would be difficult. (This problem is dealt with in a later section on exam analysis.)

Theoretical frame

The theoretical frame was developed from Alderson and Wall (1993), specifically using their fifteen washback hypotheses. The variables expressed in these hypotheses were subsequently diagrammed to conceptualise washback as a phenomenon, as shown in Figure 2.

Redefine theoretical frame

The major task at this stage was to conceptualise washback as a phenomenon by examining all the information gathered at the previous stages. First, by reconsidering the set of assumptions summarised in the practice component (see above), washback was conceptualised on the three bipolar dimensions outlined below.

1) *Specificity.* This is best illustrated by example - there is an assumption that if a test has a listening component, then the teaching of listening will be emphasized in classes. This is placed on the highly specific end of the dimension, since this assumption refers specifically to tests which include a listening component. On the other hand, there is an assumption that entrance exams will drive students to work harder than otherwise. This type of assumption does not refer to any specific type of exam, but rather to the impact of the exams in general. Thus, this sort of claim is placed on the lower end of the specificity dimension.

2) *Observability.* This dimension ranges from washback on overt behaviours to that on covert internal behaviours. If we address, say, the assumption that the impact of the exam on students leads to changes in

their reading strategies, then this type of problem is placed on the less observable end of the scale. On the other hand, if the exam affects teaching and/or learning in a way that leads teachers and students to engage in mechanical drills rather than communication activities, then this type of issue is placed on the more observable end of the dimension.

3) *Intentionality*. Intentionality may be 'direct' or 'indirect.' Direct washback occurs when the teacher has a clear intention of preparing students for exams. Indirect washback is observed when the teacher does not have such an intention, but nevertheless uses materials similar to those often included in exams. For example, there was a case in the present research where the teacher was using past exam papers without any intention of preparing students for exams. This type of washback was deemed to be 'indirect' rather than 'direct,' and was treated separately.

These dimensions proved valuable in making decisions at subsequent stages of the research. For example, the specificity dimension is implicated in selecting types of target lessons and exams. That is, to investigate the more specific assumptions, it would be necessary to select courses which aimed at preparing students for a specific university exam (out of the hundreds of different types). On the other hand, to focus on the less specific assumptions in the research, it would be necessary to select those courses which aim to prepare students for examinations generally as well as regular courses without any preparatory purpose.

The observability dimension affected the construction of research tools, since investigation of internal learning processes demands different instruments from those required for observable phenomena.

Intentionality was investigated through interviews which explored the reason behind various incidents which had taken place in the classroom.

The relationship between the variables identified by Alderson and Wall (1993) were conceptualised in a diagrammatic form in Figure 2. Note that 'micro- and 'macro-contexts' have been added as variables in the present research, since it was deemed important to spell out these variables clearly, so that the meaning of the test was clear to other researchers. Without going into further details about the diagram (see Watanabe, in progress), suffice to say that it provided a useful starting point for planning the research, although this initial conceptualisation underwent modifications as the investigation progressed.

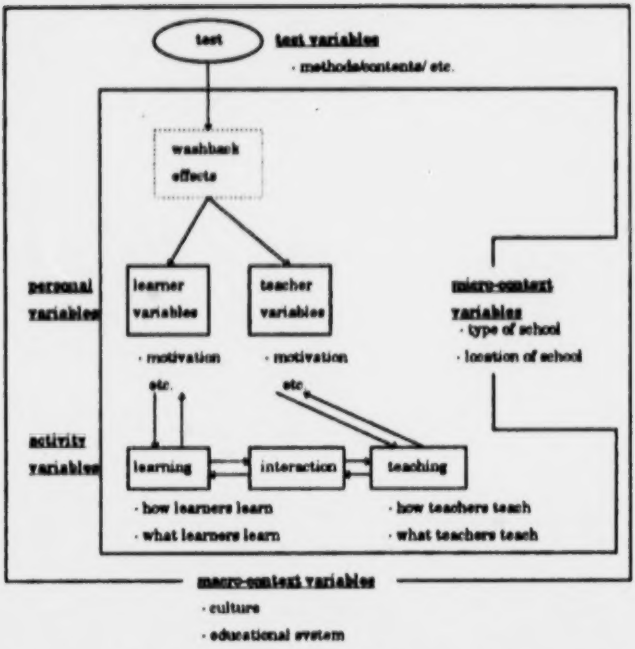


Figure 2: Relationship between variables in washback research

A place to begin research

After establishing an overall theoretical framework, the main observation research started. The task at this stage is divided into three components; formulating a working definition of the term "washback", defining the scope of the research, and determining methodology.

In the present research, washback was defined as "the impact of the examinations on what would happen in the classroom" (Alderson and Wall, 1993: 115). At this stage, the research methodology was further refined. For example, it was decided to adopt a 'cross-sectional' rather than 'longitudinal' approach; that is, in order to examine washback, different groups were observed within the same time span, rather than the same group at different points in time. It was felt that the presence or absence of washback could be satisfactorily established by comparing and contrasting the classroom events relating to a range of different exams.

Identification and analysis of the target exams

As mentioned above, in the Japanese context, where more than 1,000 different types of exams are administered every year, it was not possible to analyse all the exams. Thus, reference was made to a compilation of past exam papers published by Obunsha, a private company which deals with exam-related materials. In the 1994 edition, a total of the 280 most popular university exams were compiled, including 101 from national/ local public university exams, and 179 from private university exams. From the analysis of the papers, the following question types were considered relevant to the general assumptions (see above) of the present research:

tests measuring aural/ oral aspects of English

- listening
- speaking (indirect, using written dialogue)
- paper test of pronunciation

reading tests

- translation (from English into Japanese)
- summary in Japanese
- gap-filling
- multiple-choice questions
- questions requiring a visual response

writing tests

- direct writing tests (e.g., essay, picture description, etc.)
- translation (from Japanese into English)

tests of grammar and word usage

vocabulary tests

On the basis of these categories, the number of universities employing each question type was counted and the following tendencies were observed. There were marked differences between the exam formats of national/local public university exams on the one hand and of private university exams on the other. National/local public exams comprised

73% of 'constructive response types' (requiring testees to produce answers in their own words); with the remaining 27% being 'selective response questions' (which require students to select an answer from options). Private university exams were made up of 91% of 'selective response questions' and only 9% of 'constructive response questions'. The constructive response type questions were generally either translations (from English into Japanese or vice versa) or passage summaries.

Common tendencies emerging from the analysis were first, that only 13% of the examinations included direct listening tests; second, only 10% included direct writing tests; third, only 5% employed reading questions requiring a visual response (e.g., picture, diagram, etc.); and fourth, approximately a quarter included de-contextualised grammar, word usage, and vocabulary questions.

Prediction

Predicted research outcomes were based on results of the test analysis. These ranged from general to specific on the basis of the specificity dimension established earlier. At the most general level, it was predicted that if washback existed, there would be differences in activities and materials used in regular lessons compared to exam preparatory lessons, where both were taught by the same teacher. More specifically, it was predicted that regular lessons would be taught based on the Ministry of Education textbooks, but that exam preparatory lessons would be characterised by:

- neglect of aural/oral aspects of English (Assumption 1)
- emphasis on de-contextualised mechanical drills (Assumption 2)
- emphasis on teaching of test-taking techniques (Assumption 3)
- heavy reliance on grammar-translation (Assumption 4) and
- greater student enthusiasm for learning English (Assumption 5).

Most specifically it was predicted that if washback did exist, there would be differences in courses designed for different target university exams. These differences were expected to manifest in the following areas:

- test method; i.e., objective vs. subjective, direct vs. indirect (Assumption 6)
- the skill area, listening in particular (Assumption 7), and
- grammar-translation (Assumption 8).

Selection[1]

Prediction 1 was to be investigated by observing high school classes. It was expected that teaching in regular lessons would reflect the communicative orientation expressed in the Ministry of Education guidelines, while the exam preparatory lessons would be oriented towards more formal aspects of English reflecting the exams. However, high schools were not appropriate for attesting the more specific predictions 2 and 3 since high school students vary widely in terms of their target exams, so the teacher must teach students aiming for various target exams at the same time. On the other hand, private extra-curricular institutions, (*yobiko* literally, 'preparatory schools') were ideal, since they offer a range of courses targeting specific university exams.

Schools were selected based on the following criteria: (i) in high schools, two teachers were selected who taught the same preparatory and regular courses; (ii) in the *yobiko*, two teachers were selected who taught two different target university exam courses; i.e., both taught courses aimed at the exam for national/ local public universities and the private university exams. Both teachers selected taught to the same two exams.

Access and entry

In identifying the schools, an attempt was made to select those with a reputation for successful student acceptance at prestigious universities, since it was assumed washback would be observable in such schools. However, various school background variables were also considered in the selection, such as educational methods, curricula, student characteristics, etc., since it was conceivable that if exams exercised sufficient power over education, then washback would override these background variables.

The following observations were made:

High School A			
Teacher A1	Regular	79 minutes	June 24, '94
	Exam preparation	79 minutes	June 27, '94
Teacher A2	Regular	57 minutes	June 25, '94
	Exam preparation	67 minutes	June 24, '94
High School B			
Teacher B1	Regular	98 minutes	June 29, '94
	Exam preparation	80 minutes	June 30, '94
High School C			
Teacher C1	Regular	47 minutes	Sept 29, '94
	Exam preparation	47 minutes	Oct 6, '94
Teacher C2	Regular	50 minutes	Nov 17, '94
	Exam preparation	48 minutes	Nov 28, '94
Yobiko			
Teacher Y1	A private university course	456 minutes	Aug 4 to 9, '94
	A national university course	451 minutes	Aug 4 to 9, '94
Teacher Y2	A private university course	508 minutes	Dec 25 to 29, '94
	A national university course	382 minutes	Jan 3 to 7, '94

Note that practical constraints meant that the selected samples were not ideal. First, in High School B it was not possible to achieve a perfect "pairing", with two different teachers teaching both regular and exam preparatory lessons, Teacher B1 was in fact the only teacher who taught both types of course. Also the length of observations possible in high schools was limited because of a conflict of schedules - the researcher could only carry out observations during the intersessions between terms at university. Unfortunately, high schools were also on holidays during these periods. In addition, high schools had various extra activities, such

as mid-term exams, final exams, sport festivals, cultural festivals, etc. which made regular observation difficult. On the other hand, with the *yobiko* it was possible to observe whole intensive courses for specific target university exams, which were held during summer and winter holidays.

High School A & B were public schools located in rural areas. High School A had quite a high reputation for successful preparatory teaching in the area. High School B was reputed both for exam preparation and for their unique English teaching method - a type of immersion program where the entire lesson was taught in English. High School C was a private school located in a metropolitan area. This high school attracted numerous returning students, who had graduated from junior high schools in English-speaking countries. The *yobiko* was located also in a metropolitan area. This school attracted students from all over Japan, and had a high reputation for its emphasis on teaching English for entrance examinations.

Data collection and analysis instruments

While access to schools was being established, the data collection instruments were constructed. There were two major components.

The first comprised sets of pre-observation interview questions. Questions focused on information about the teacher's background, for example, education, age, major field of study. Questions about the teacher's perception of the entrance exam were avoided since such questions might affect his/her teaching, which in turn might pollute observations.

The second component consisted of lesson description sheets (in addition to pencils, audio-recorder, stop-watch). The observation sheets were designed to record the following types of information: time sequence, materials used, teacher activities/speech, student activities/speech, what was written on the chalkboard, observer's comments and the questions which came to the observer during the observation. These questions were asked of the teacher at post-observation interviews.

Enter the classroom

Non-participant observation was employed in the research: that is, the observer's involvement was kept to a minimum and the observer made

every attempt to avoid disturbing the lesson. When the teacher entered the room, stop-watch and audio-recording were started at the same time.

The initial purpose of the observation was to develop 'low inference' categories - clear enough to be specified in behavioural terms. An attempt was made to record new events, felt to be relevant to the research predictions, in as precise and concrete a manner as possible.

Each lesson was recorded on audio tape, to permit subsequent frequency counts based on the categories developed.

Post-observation interviews

After observing the lessons, post-observation interviews were conducted with the teachers concerned. The interview had two major purposes: to gather data about the teacher's perception of the impact of the exams, and to examine the intentions which lay behind the various 'events' occurring in the classroom. These data served as a source for interpreting the results.

Create data sources

Immediately following the observation and interviews, while memory was still fresh, the lesson description sheets were reviewed, and ambiguous parts clarified by listening to the audio tape. In addition, the interview data was summarised in propositional formats. Subsequently, the descriptions of the classroom events were summarised in the form of 'narrative vignettes', ("a more elaborated, literally polished version of the account found in the field-notes" Erickson, 1986: 150). The vignettes were designed to illustrate the overall picture of the classrooms and to assist with remembering the overall flow of classroom discourse at the stage of the frequency analysis.

At this stage, the classroom materials, which consisted of the textbook and supplementary exercises, were analysed according to the question type categories identified at the stage of the exam analysis (see above).

Data analysis

Both materials and classroom events were analysed. The latter are focussed upon here as these were the core of the research process. The

purpose of the data analysis was two-fold: first to identify categories relevant to the predictions, and systematise them into a coding scheme for counting the frequency of incidents; and second, to code incidents which belonged to each category by listening to the audio tape.

Where possible, the data analysis was conducted immediately after the observation. The initial stage of the analysis involved repeated scannings of the lesson description sheets in order to become familiar with the data. Next, various incidents deemed relevant to the research predictions were circled. The identified incidents were word-processed in propositional form, and labels were attached indicating the nature of the incident, the teacher involved and the context in which the event occurred. For example:

23/6/1994

1 HA/TAe teacher reads aloud a paragraph.

t's turn/ read aloud/ a paragraph/ aural-oral/ English

24/6/1994

2 HA/TAr teacher calls on students to translate a sentence.

t's turn/ interact/ elicit/ Jpn/ a sentence

24/6/1994

3 HA/TAr student translates English into Japanese.

s's turn/ translate / a sentence/ Jpn —> Engl.

HA = high school A
TA = Teacher A
e = exam preparatory lesson
r = regular lesson

Incidents were dated and numbered for future reference. After identifying and labeling all the incidents, they were classified into categories. To systematise the categories, reference was made to the COLT (Communicative Orientation of Language Teaching) scheme

(Allen, Froehlich and Spada 1984) because of its flexibility and comprehensiveness for capturing a wide range of communicative features of language teaching. This system was referred to in two ways; first, some category labels of the system were adapted to the categories derived from the analysis of the field-notes; and second, the format of the system was employed to systematise the separate categories into a larger system and render it workable for actual data coding.

All the derived categories were then systematised into a typology as follows:

- classroom organisation patterns
  - teacher-fronted
  - pair-work
  - group-work
  - students' presentation
- interaction
  - turn-taker
    - teacher or student/s
  - language medium
    - English or Japanese
  - types of information and exchange patterns
    - requesting pseudo or genuine information
    - giving predictable or unpredictable information
    - reacting to form or message
- activities
  - choral
  - read aloud
  - paraphrase
  - translation from English into Japanese
  - translation from Japanese into English
  - summary in Japanese
  - summary in English

topics

examination
language form or use

use of metalanguage

incidents of laughter

The last category 'incidents of laughter' had not been predicted, but the category was incorporated, since it was found during observations that laughter was an important indicator of classroom atmosphere (cf., Alderson and Hamp-Lyons, 1994).

On the basis of the typology, a new coding scheme, COEPREC (Communicative Orientation for Exam Preparatory Classes), was developed. A sample copy is provided in Appendix A. Like COLT, the system is divided into two sheets, A and B. Sheet A was prepared to measure 'classroom organisation patterns' and 'the use of materials', which were appropriately measured by referring to the length of time, while Sheet B, various categories, which were to be measured by 'frequency'. The audio-recording was played again to code classroom events on COEPREC, and the length of time and the frequency were subsequently computed. Because the focus of this report is on the qualitative aspects of the research, we will not elaborate here on the coding and the data analysis procedures. Interested readers are referred to Watanabe (in progress).

Integration

The results of the coding were subsequently summarised into tables, called 'Washback Grid.' A sample table, which contains the information concerning teaching methodology of high schools, is provided in Appendix B. Note that when summarising data into tables, some COEPREC categories were combined with others. For example, in order to examine the frequency of using English aurally/orally, all the utterances made in English for interaction and activity were combined into one larger category "aural/oral." By examining closely the whole table, three forms of washback were observed in each category; i) washback does not exist, ii) washback exists with some teachers but not

with others, and iii) washback exists with all teachers. For example, the table in Appendix indicates that the examinations influenced some teachers but not other teachers in the area of 'types of information exchanged'. It had been predicted from Assumption 2 (i.e., in exam preparatory lessons, a greater emphasis is placed on de-contexualised mechanical drills than communicative exercises) that 'pseudo/ predictable' information, (i.e., the type of information that is known between the two persons, thus, indicating that the mechanical exercises were conducted) would be exchanged more frequently in exam preparatory than in regular lessons. However, in results obtained thus far, this type of assumption was not completely supported; washback was clearly present only in the lessons of Teacher C, whereas in the lessons of other teachers evidence of washback was slight.

Review

After the data had been summarised, the results were interpreted by referring to the pre-/post-observation interview data. This task was important in answering the basic question posed earlier: whether washback does or does not exist. From time to time during the process of interpreting the classroom events against the interview data, several possible interpretations emerged. When this occurred each of the "rival" explanations was noted.

For example, listening was taught in some of the high school exam preparatory lessons, but not in others. One of the reasons emerging from interviews was that the teachers who included listening in their class believed that the exam preparation could provide students with a good opportunity for developing communication skills. On the other hand, those who did not include listening claimed that there were too few support-materials available, and that they did not know how to teach listening skills effectively.

Another example is the use of the grammar-translation method. It has generally been argued that the teaching methods of pre-college EFL in Japan are based on grammar-translation because the university entrance examinations require translation and grammatical analyses. However, it seemed that teacher orientation was a more important factor in determining what happened in the classroom than the content of the

exam. One teacher used grammar-translation irrespective of the target exam, while the other used this method when he was explaining grammatically complex sentences, and when he was teaching for the target exam which often tested translation and complex grammatical structures. Thus, the latter teacher seemed to be influenced by the exam, while the former did not. The differences seem to be partly attributable to the teachers' educational backgrounds. The teacher who used grammar-translation had majored in theoretical linguistics at post-graduate level, while the other teacher who translated texts only occasionally, had gone through a teacher training course where communicative teaching was emphasised. Another possible source of the difference seemed to come from the teachers' past experience of learning English; for example one teacher reported that he felt it very difficult to teach English by other methods than the one by which he was taught.

At this stage, in addition to interpreting the results, the whole research process was also reviewed by examining what was lacking and what could be done in future research. For instance, as mentioned earlier, this was a cross-sectional study, so it left unanswered the question of whether teaching strategies would change as the exam period approached. Such an issue can only be addressed through longitudinal research, in which the same class is observed at different periods over time.

Drawing implications

The research findings thus far suggest that the incidents which were observed in each lesson did not greatly differ according to the purpose of the lessons (i.e., regular or exam preparatory) of high schools or the target exams of the *yobiko* lessons (i.e., national/local public or private). Differences between teachers and schools were great. The results suggest that changes to exams will not automatically lead to changes in teaching; but rather, teacher factors seem to play a greater role in deciding what happens in the classroom. It may be that in-service/pre-service teacher training should incorporate teaching methods which show how exams can be used to develop communicative ability. As the research progresses, however, it is expected that other implications will be drawn for other audiences, including test constructors, the Ministry of Education, etc.

CONCLUSION

In this paper, an approach to washback research which draws on ethnographic methods has been described. Despite the initial intention of demystifying the research procedure, the description may still seem to be highly complex. Such complexity is necessary and desirable because the phenomenon of washback itself is complex, and we are still at a preliminary stage of the research.

I would argue that an approach informed by the principles of ethnographic research is an appropriate way to deepen our understanding of the nature of washback. In support of this argument I now return briefly to the four research strategies identified by LeCompte and Preissle (1993) which were referred to earlier in this paper.

1) Ethnography elicits phenomenological data that represent the world view of the participants being investigated and participants' constructs are used to structure the research.

In ethnographic research, a researcher is only one of a range of participants, so his/her perception represents only view amongst many. Because tests are used in a particular context for a specific purpose, it is important before embarking on the research to identify the problems which participants themselves see as important. Otherwise, the research may neglect problems considered crucial by participants in favour of those which may be deemed irrelevant and extraneous. Consequently, the research results are likely to be sterile, having few implications for those most affected by the problems. It is for this reason that I have attempted to base my research not only on the findings of previous research, and on my own experience as an EFL teacher but also on views of washback gathered from a range of different sources, including other teachers, students and the media.

2) Ethnography employs participant and non-participant observation to acquire firsthand, sensory accounts of phenomena as they occur in real world settings.

There are two key phrases in this statement; "firsthand, sensory accounts of phenomena" and "real world settings." First, if the washback research were not to gather firsthand data, we would be obliged to take what teachers and students say about how they feel

about the effects of examinations at face value, even though their perceptions may not reflect what they are actually doing. As Hopkins correctly points out

> there is often incongruence between a teacher's publicly declared philosophy or beliefs about education and how he or she behaves in the classroom.
>
> (Hopkins, 1985: 48)

Second, ethnographic research stresses gathering natural data. If for example it were to be found that a test administered as part of a controlled experiment had no impact on teaching and learning, this result could not be generalized to actual exam preparatory courses since tests administered under experimental conditions would be likely to be perceived by participants as having no educational consequences. Tests always play a certain role in a specific context, so washback research needs to take these contextual factors into account.

3) In ethnographic research, the researchers seek to construct descriptions of total phenomena within their various contexts and to generate from these descriptions the complex interrelationship of causes and consequences that affect human behaviour toward and belief about phenomena.

As pointed out by Alderson & Wall (1993), the exam may be only one of the factors that affect how innovations succeed or fail, and this study also seems to support this assertion. In other words, numerous variables other than the exam, such as the teacher's beliefs, his/her familiarity with the teaching method, etc., seem to be involved in determining what happens in the classroom. This type of insight could not have been gained without an attempt to describe the total phenomena of the classroom, including the teacher's perceptions about his/her teaching.

The small portions of the on-going research reported in this paper are sufficient to indicate that the washback effects are highly complex phenomena - a fact which is somewhat at odds with popular opinion. Educational innovations do not seem to spring automatically from innovations to examinations. Rohlen, an anthropologist, once observed five Japanese high schools, and

found that there was a large amount of variation among them. He states:

> Neither Japanese society nor its high schools are monolithic, a point too regularly ignored in foreign treatments of Japan.
>
> (Rohlen, 1983: 43)

His final comment is also suggestive for washback research;

> It is easy to suggest cause and effect relations between schooling and the lamentable aspects of society, and it is equally difficult to disprove them. Sweeping interpretations containing some answers appeal to the popular mind and are useful in campaigns to reform education.
>
> (Rohlen, 1983: 327)

The present research likewise warns us against basing educational innovation solely on public opinion.

4) Ethnographic researchers use a variety of research techniques to amass their data.

Direct observation is important, but not the only method appropriate for washback research. Rather, various research methods, including self-report, document analysis and interview, should be considered to complement each other. Without the interviews, for example, it would not have been possible to detect the reasons (or intentions) behind teachers' behaviour in the classroom. The problem is which method should be employed at which point, and it is hoped that this paper will help researchers to make an informed decision in conducting their own research.

Before establishing a theory of washback, we clearly need more empirical data gathered from different contexts. Results of such research will be of great assistance to all those involved in testing, when they have to make decisions about the construction and use of tests. Pilliner (1973) once said: "the most important requirement of a good exam is that it should be educationally beneficial" but, regrettably, this has thus far been a neglected aspect in the cycle of testing.

NOTES

1 'Selection' is the qualitative equivalent to the term 'sampling' used in the positivistic tradition. According to LeCompte and Preissle (1993), "selection requires only that the researcher delineate precisely the relevant population or phenomenon for investigating, using criteria based on theoretical or conceptual considerations, personal curiosity, empirical characteristics, or some other consideration" (p. 57).

REFERENCES

Alderson, J. C. and D. Wall (1993) Does washback exist? *Applied Linguistics* 14,2:115-129.

Alderson, J. C. and L. Hamp-Lyons (1994) TOEFL preparation and the communicative classroom. Paper presented at TESOL conference, Baltimore, March.

Allen, P., M. Froehlich, and N. Spada (1984) The communicative orientation of language teaching: and observation scheme. In J. Handscombe, R.A. Orem, and B. P.Tayler (eds) *On TESOL '83*. Washington D.C: TESOL.

Ariyoshi, H., and K. Senba (1983) Daigaku nyushi junbi kyoiku ni kansuru kenkyu (A study on preparatory teaching for the university entrance examinations). *Fukuoka Kyoiku Daigaku Kiyo*, 33: 1-21.

Berwick, R. and S. Ross (1989) Motivation after matriculation: are Japanese learners of English still alive after exam hell? *JALT Journal*, 11: 193-210.

Erickson, F. (1986) Qualitative methods in research on teaching. In Wittrock, M.C. (ed.)*The handbook of research in teaching*, 3rd ed. New York, Macmillan.

Evertson, C. and J. Green (1986) Observations as inquiry and method. In Wittrock, M.C. (ed.) (1986) *The handbook of research in teaching*, 3rd ed. New York, Macmillan.

Fujita, T. (1992) Readiness model of optimal input: a comparison between Japanese high school and non-high school students of English. *Sophia Linguistica*, 31: 122-143.

Hopkins, D. (1985) *A teacher's guide to classroom research.* Milton Keynes, Open University Press.

LeCompte, M. D. and J. Preissle (1993) *Ethnography and qualitative design in educational research.* San Diego, Academic Press.

Ministry of Education (1994) *Monbu Toke Yoran* (Statistical abstract of education, science and culture). Tokyo, The Ministry of Education.

Obunsha (1994) *Zenkoku daigaku nyushi mondai seikai* (Compilation of past exam papers). Tokyo, Obunsha.

Pilliner, A. (1973) Assessment - principles and practice with special reference to education in Pakistan. Unpublished manuscript, The British Council.

Rohlen, T. P. (1983) *Japan's high schools.* Berkeley,University of Berkeley Press.

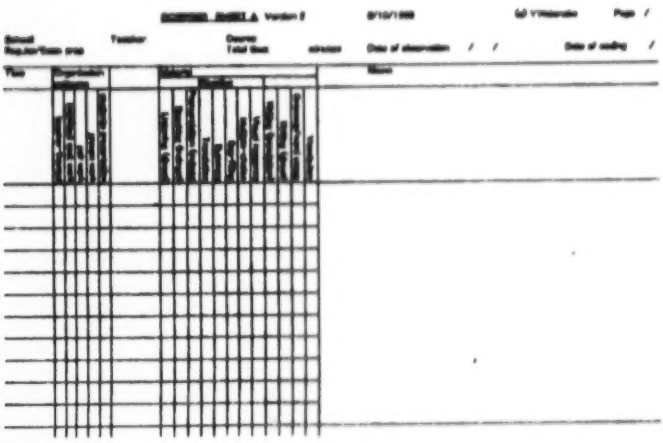
Saito, T., S. Arita, and I., Nasu (1984) Taki-sentaku tesuto ga igaku kyoiku ni oyobosu eikyo (Effects of multiple-choice questions on medical education). *Nihon igaku kyoiku shinko zaidan kenkyu jose ni yoru kenkyu hokoku sho.*

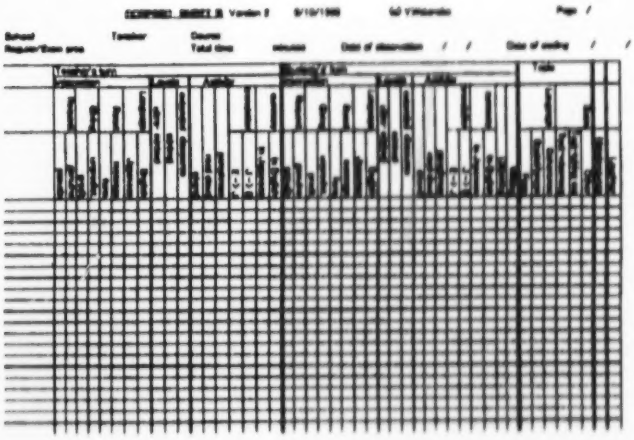
Wall, D. and J.C.Alderson (1993) Examining washback: the Sri Lankan impact study. *Language Testing*, 10, 1: 41-69.

Watanabe, Y. (1990) External variables affecting language learning strategies of Japanese EFL learners. MA dissertation, Lancaster University. (ERIC Document Reproduction Service No. ED 334 822).

Watanabe, Y. (in progress) Washback effects of Japanese university entrance examinations - classroom-based research (a tentative title). Ph.D. thesis. Lancaster University.

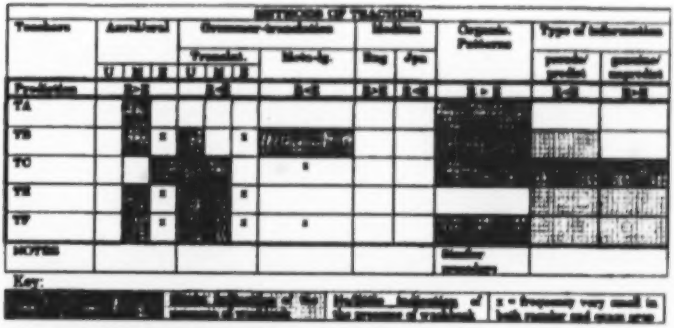
APPENDIX A: CODING SCHEME (COPREC)





BEST COPY AVAILABLE

APPENDIX 2: WASHBACK GRID



244

BEST COPY AVAILABLE

AUTHOR NOTES

Rosemary Baker is a National Health and Medical Research Council Research Officer in the Department of Speech Pathology and Audiology at the University of Queensland. Her major fields of research are the communication abilities of the ethnic aged and the assessment of language impairment resulting from stroke and Alzheimer's dementia in bilinguals and speakers of English as a second language. She also works on the development of language proficiency tests.

Annie Brown is a Research Fellow within the National Languages and Literacy Institute of Australia Language Testing Research Centre. She is currently undertaking doctoral studies in the assessment of oral communication skills in occupational contexts.

Micheline Chalhoub-Deville is a Research Fellow in the Center for Advanced Research on Language Acquisition (CARLA) and a Graduate Faculty member in the ESL program at the University of Minnesota. She is currently working on a computer adaptive testing (CAT) project for assessing reading proficiency in French, German and Spanish. Dr. Chalhoub-Deville has authored several articles, reports and presentations on performance assessment, reading comprehension, test evaluation and CAT. She has also directed the development of several university placement and exit proficiency tests. She has been elected to the Executive Board of the American Association of Teachers of Arabic (AATA). She is also a member-at-large of the Test of English as a Foreign Language (TOEFL) Research Committee.

Catherine Elder has taught Italian in a number of university departments and since 1990 has held a position as Research Fellow in the NLLIA Language Testing Research Centre (LTRC) at the University of Melbourne where she has worked on a variety of projects involving test development and research. She has an MA in Applied Linguistics from the University of Melbourne and is currently completing a PhD in the area of language testing. Her research interests are in language testing, bilingualism, classroom-based foreign/second language learning and language policy.

Lesleyanne Hawthorne is the author of *Labour Market Barriers for Immigrant Engineers in Australia* (AGPS 1994), *Making it in Australia* (Edward Arnold 1988) and *Refugee - The Vietnamese Experience* (OUP 1982), as well as co-author of *English for Occupational Purposes* (RMIT 1988). She holds the position of Research Manager at the Bureau of Immigration, Multicultural and Population Research, where she manages research related to labour market, language and health issues, and also lectures in Languages for Specific Purposes at Monash University. Lesleyanne is currently undertaking a major study of immigrant nurses in Australia, including analysis of the demand for bilingual bicultural nursing skills.

Tom Lumley is a Research Fellow within the National Languages and Literacy Institute of Australia Language Testing Research Centre. He is currently undertaking doctoral studies, in the assessment of writing skills.

Joy McQueen is a Research Fellow at the Australian Council for Educational Research. She has been a teacher and administrator of LOTE and ESL programmes in primary, secondary and tertiary settings, and has worked in educational research and assessment at the NLLIA Language Testing Centre (University of Melbourne) and at ACER.

John Read is Senior Lecturer in Applied Linguistics at Victoria University of Wellington. His major research interests are in language testing, second language vocabulary knowledge and the sociology of language. He has published a number of articles in these areas and is currently writing a book on vocabulary testing.

Nikhat Shameem is a Lecturer in applied linguistics at the University of Auckland. Her interests are in sociolinguistics and language assessment, particularly in the area of minority languages, and in English language teaching.

Carolyn E. Turner is Associate Professor and Certificate Program Director in the Department of Second Language Education at McGill University, Montreal. Her major fields of research are testing and

evaluation and curriculum development in second language education. She is co-author with Denise Lussier of *Le Point sur...L'évaluation en didactique des langues* (Second Language Evaluation: The State of the Art), Centre Éducatif et Culturel Inc., 1995.

John A. Upshur is Professor of Applied Linguistics at Concordia University where he teaches courses in language testing and research methods. He has worked in six different countries developing ESL tests. Dr. Upshur served as editor of the journal Language Learning and, more recently, of Language Testing.

Yoshinori Watanabe is a Lecturer in the English Language Program at the International Christian University. His major fields of research include language learning strategies, and effects of exams on teaching and learning of English as a foreign language. He is currently undertaking research for a doctoral degree at Lancaster University.

Gillian Wigglesworth is a Lecturer in the Department of Linguistics at Macquarie University. She completed her PhD in Linguistics 1992 in the area of first language acquisition, and then worked at the University of Melbourne in the Departments of Linguistics and Applied Linguistics and at the NLLIA Language Testing Research Centre. Her research interests include first and second language acquisition, language assessment and evaluation and bilingualism.